DOCTORAL THESIS

Identification of single nucleotide polymorphisms (SNPs) involved in the determination of craniofacial morphology

Barash, Mark

Award date:
2014

BOND UNIVERSITY
FACULTY OF HEALTH SCIENCES & MEDICINE

Identification of single nucleotide polymorphisms (SNPs) involved in the determination of craniofacial morphology

Mark Barash

Faculty of Health Sciences and Medicine

Bond University

A thesis submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy by Research

Abstract

Current methods of forensic DNA analysis are mainly used for identification purposes and require a reference sample for comparison to an evidence DNA profile. The short tandem repeats (STR) is presently the major highly discriminating method for forensic DNA analysis; however, it does not provide useful investigative information, with an exception of sex. The ability to provide information on visual appearance of a person has great investigative potential in various types of criminal and mass disaster investigations. In the last few years this rationale has given rise to a new discipline of Forensic Molecular Phenotyping. To date, most work in this area has concentrated on pigmentation traits (eye, skin and hair colour), as well as on bio-geographic ancestry. Several forensic assays were developed, which were able to predict some of the eye and hair shades, as well as several ancestries. However, the genetics of the most intriguing part of the human appearance – the face, has not been explored extensively.

This project aimed to identify single nucleotide polymorphisms (SNPs) influencing craniofacial morphology and subsequently incorporate a set of significantly associated markers in forensic molecular Identikit for prediction of facial appearance. The work to achieve this goal included an implementation of a candidate gene approach along with collecting almost 600 DNA samples, accompanied with 3 dimensional (3D) facial images. Specifically, more than 1,200 genetic markers located in genes, potentially involved in craniofacial embryonic development were genotyped using a Massively Parallel Sequencing (MPS) platform, while the 3D images were analysed for 92 various craniofacial measurements and ratios.

In order to validate the statistical methods used for genetic association analysis, a set of markers, previously associated with pigmentation and ancestry were incorporated in the sequencing panel. Additional markers, such as autosomal STRs, INDELS and identity-informative SNPs were included in the final genotyping panel as well. The final custom amplicon set included 1,700 amplicons, covering approximately 6,500 markers with various minor allele frequency, which were genotyped in 587 DNA samples.

The statistical analysis was performed under stringent conditions and identified multiple

associations between craniofacial traits and candidate genetic markers. Over 160 SNPs in 63 genes and intergenic regions were strongly associated with either 3D craniofacial measurements or principal components, representing facial shapes (areas). Most of these genes were previously associated with either embryonic development or craniofacial syndromes, although for a few of them, this link was not described before. The association analysis of the pigmentation traits revealed multiple associations with both known pigmentation genes and novel genes, not yet linked to pigmentation. The analysis of four main population groups showed strong association with numerous markers, mostly associated with pigmentation, as well as novel ancestry-informative SNPs. Ancestry and pigmentation association results were in consensus with published data and provided verification of statistical methods used for analysis of the craniofacial traits.

The association results provided strong evidence that novel polymorphisms are involved in the normal variation of craniofacial morphology. Subsequently, these results enabled creation of statistical models for potential prediction of craniofacial traits along with pigmentation and ancestry.

Declaration

This thesis is submitted to Bond University in fulfilment of the requirements for the degree of Doctor of Philosophy by Research.

This thesis represents my own original work toward this research degree and contain no material which has been previously submitted for a degree or diploma at this University or any other institution, except where due acknowledgment is made (Sections 3.5.1 and 3.5.2).

Mark Barash

March 2014

Acknowledgements

First and foremost, I would like to thank my family for their extensive support and patience during these four long years. It has been not easy for my wife and children to relocate to Australia and our families to stay back in Israel for such a long time.

I would also like to sincerely thank all the volunteers who participated in this project. Without their participation and support, this study could not have been accomplished.

I would like to sincerely thank my supervisors Professor Angela van Daal and Associate Professor Lotti Tajouri for all the support, advice and continuous assistance during this project. I came to Bond University and continued with my studies despite all the circumstances only, because I wanted to do a research project under Angela's supervision.

I would like to acknowledge the Health Science and Medicine Faculty of Bond University for providing me with a stipend for the first year of my project as well as sufficient funds to present its outcomes at several national and international conferences. I must also thank the Australian Government for providing me with two and a half years International Postgraduate Research Stipend and the Pelerman Holdings Pt Ltd. for another six months stipend. Additional funding for the consumables used in this research was provided by the United States National Institute of Justice (NIJ), and the Technical Support Working Group (TSWG) of the United States Government National Interagency Research and Development Program for Combating Terrorism. In addition, substantial support was provided by Illumina, Inc. in the form of long-term loan of the BeadExpress™ scanner, custom primer design, and the provision of all consumables. I definitely could not succeed in this journey without this financial support.

Many thanks goes to my fellow HDR students from our small forensic team, namely Sheree Hughes-Stamm, Kelly Grisedale, Olga Kondrashova and Kat Sanders who helped and supported me all the way through my PhD.

On a personal note, I would like to especially thank my dear friend and colleague Ayeleth Reshef from Israel Police. She is my true mentor and the first person who taught me the essentials of the forensic DNA analysis. She believed in me and helped with all the means to make my dream come true.

And finally, I would like to thank my grandmother Rachel and grandfather Alexander (ז"ל), whose influence on my personality was tremendous. This work is dedicated to you!

Table of Contents

List of Terms and Abbreviations

Anthropometrical terms

Alare (al): Instrumentally determined as the most lateral points on the nasal aperture in a transverse plane.

Anencephaly: The cephalic malformation originating from a neural tube defect that occurs when the cephalic end of the neural tube fails to close. Usually between the 23rd and 26th day of pregnancy, resulting in the absence of a major portion of the brain, skull, and scalp.

Bizygomatic Diameter (zy-zy): Direct distance between most lateral points on the zygomatic arches (zy-zy).

Cephalometry: Scientific measurements of the craniofacial bones.

Euryon (eu): The most laterally positioned point on the side of the braincase (paired). Euryon always falls on either the parietal bone or on the upper portion of the temporal bone and may be determined only by measuring maximum cranial breadth.

Exencephaly: A type of cephalic disorder wherein the brain is located outside of the skull.

Frankfurt position (auriculo-orbital plane): Anatomical position of the human skull close to the position of the head normally carried by the living subject.

Glabella (g) (nasal eminence): The most forwardly projecting point in the mid-sagittal plane at the lower margin of the frontal bone, which lies above the nasal root and between the superciliary arches.

Gnathion (gn): The lowest point on the lower border of the chin.

Gonion (go): A point along the rounded posterior corner of the mandible between the ramus and the body (paired).

Holoprosencephaly (HPE): Cephalic disorder in which the prosencephalon (the forebrain of the embryo) fails to develop into two hemispheres.

Maximum Cranial Length (g-op): Distance between glabella (g) and opisthocranion (op) in the midsagittal plane, measured in a straight line.

Nasion (n): The point of intersection between the frontonasal suture and the midsagittal plane.

Neural crest: Embryonic tissue, consisting of multipotent migratory cell population that gives rise to a diverse cell lineages including melanocytes, craniofacial cartilage and bone, smooth muscle, neurons and glia.

Neural tube: In vertebrates, this is the embryo's precursor to the central nervous system, which comprises the brain and spinal cord.

Opisthocranion (op) (opisthocranium): The most posteriorly protruding point on the back of the braincase, located in the mid-sagittal plane. Opisthocranion almost always falls on the superior squama of the occipital bone, and only occasionally on the external occipital protuberance.

Polygon mesh: A 3D representation of an object consisting of an ordered set of three or more points called vertices, each vertex connected by line segments called sides or edges to two other vertices.

Vertex (v): The highest point of the head when positioned in the Frankfort Horizontal Plane.

Zygion (zy): The most laterally positioned point on the zygomatic arches. The position of zygion is defined from the measurement of bizygomatic breadth.

Population genetics and statistics terms

EVCs: Externally visible characteristics. A number of phenotypic traits, such as eye, skin and hair pigmentation and facial appearance, which may have forensic relevance.

GWAS: Genome–wide association study. A genetic study that attempts to identify commonly occurring genetic variants that contribute to disease risk or phenotypic traits.

Fst: The fixation index statistic is a measure of how populations differ genetically. Fst = $(H_T-H_S)/H_T$, in which H_T and H_S represent heterozygosity of the total population and subpopulation respectively. The value of Fst can theoretically range from 0.0 (meaning no differentiation) to 1.0 (meaning complete differentiation).

Haplotype: A set of alleles, located closely together on the same chromosome, which tend to be inherited together.

iHS: Statistical test which is applied to individual SNPs and its large positive and negative values indicate unusually long haplotypes carrying the ancestral and derived allele, respectively.

INDELs: Insertion or deletion polymorphisms in the DNA sequence.

Linkage disequilibrium (LD): The non-random association of alleles between two or more loci.

Long range haplotype (LRH) test: This test examines the association between allele frequency and the extent of LD.

Ortholog: Genes that evolved from a common ancestral gene in different organisms and retain the same function through evolution. Used for prediction of gene function in novel genomes.

Paralog: A duplicate of a gene within the same genome. May have a different function, although usually related to the original gene.

Population stratification: The false association of an allele to studied trait due to both differences in population frequency of the allele and differences in ethnic prevalence or sampling of affected individuals.

Principal component: A composite variable that summarizes the variation across a large number of variables.

SNP: A single nucleotide polymorphism in the DNA sequence.

STR: A short tandem repeat of two or more nucleotides repeated as sequence blocks, adjacent to each other.

Tag SNP: A markers chosen from a large set of available SNPs, based on favourable linkage disequilibrium with other multiple markers.

XP-EHH: Is a cross population extended haplotype homozygosity test designed to detect ongoing or nearly fixed selective sweeps by comparing haplotypes from two populations.

List of Figures

List of Tables

Addendum

Contributions to this document:

Mark Barash:

- Experimental design
- Laboratory bench work
- Data analysis, except for generating prediction models
- Author of the document (Chapters 1-6)

Philipp Bayer

- Validation of SVS statistical analysis with PLINK software (Chapter 5)
- Generation of prediction models (Section 5.4)

Dr. Sheree Hughes – Stamm

- Laboratory assistance with GoldenGate assay (sections 3.5.1 and 3.5.2)

Prof Angela van Daal

- Assistance with experimental design and interpretation of results
- Editing of the document (Chapters 1-6)

Assoc. Prof Lotti Tajouri

- Editing of the document (Chapters 1-6)

Assoc. Prof Kevin Ashton

- Editing of the document (Chapters 1-6)

Chapter 1

Introduction and Literature Review

1.1. General Introduction

The current use of human DNA for forensic identification relies on comparison of a Short Tandem Repeat (STR) profile, from a crime scene with a known reference from a suspect or victim. If the suspect is not available, a DNA profile can be compared to a DNA database. However, in cases where no suspect has been identified and a criminal DNA database comparison has not revealed a match, the conventional STR profile from a crime scene item remains a 'black box', unable to provide information that may serve as an investigative lead (with the exception of sex). Similarly, in a mass disaster or missing person cases, a DNA profile obtained from an unknown person has to be compared with a reference sample from personal belongings or known relatives of the missing person. In some cases, a craniofacial reconstruction is performed, providing additional information on the unknown person. However, this tool is not always available and when available, not always accurate.

The ability to determine the biogeographic ancestry and visual appearance of an individual from unidentified skeletal remains or a biological specimen at a crime scene can provide important probative information to law enforcement investigators. This approach has recently become known in forensics as Forensic Molecular Phenotyping (FMP) [2, 3]. Additional information on external visible traits (EVTs), such as pigmentation, height, ancestry, and facial appearance may help to reduce the number of potential suspects and facilitate police efforts to find the perpetrator of a crime. In Disaster Victim Identification (DVI) cases, including unidentified skeletal remains, this information can also be helpful in reconstructing the visual appearance of an unknown individual.

Another benefit of using this reverse facial recognition approach is that human eyewitnesses have been shown to be inaccurate or simply mistaken [4-7]. Thus, a scientific approach, supported by statistics, would be a preferable option in some cases. Accurate prediction of EVCs from a DNA sample may help to verify eyewitness testimony or provide additional information in the absence of such.

While some physical characteristics are significantly affected by the environment, others are largely determined by the genome. These include the major physical descriptors of a person: eye, hair and skin colour, as well as height and facial features. These traits are known to be polygenic and are likely to be determined by a number of Single Nucleotide Polymorphisms (SNPs) that act in specific combinations to produce unique human phenotypes [8]. An extensive amount of work has been done in the last

few years on finding the genetic basis of pigmentation traits (eye, skin and hair colour) and detection of ancestry informative markers, which has led to the development of assays, capable of predicting these traits [9-14]. Detecting a set of polymorphisms in candidate genes responsible for the variance in craniofacial morphology will lead to the development of a robust, phenotypically informative forensic assay. This science-based "molecular identikit" will be less prone to the potential bias of an eyewitness and provide a valuable supplemental tool for solving crimes. In DVI cases, this information will help with traditional anthropological facial reconstruction by providing additional information on the soft facial tissue, which currently is reconstructed based on a best guess (as there is no accurate correlation between cranial bones shape and soft tissue organs). Finding the genetic basis for normal variety in craniofacial morphology will also extend our knowledge on understanding of the craniofacial embryonic development.

In order to be able to predict human facial appearance from DNA, it is essential to understand the basic elements of the craniofacial anthropometry and genetics background, collect a range of accurate craniofacial measurements from a large group of individuals and genotype a set of candidate genetic markers, which may be subsequently associated with specific anthropometric measurements.

1.2. Anthropometry of the head and face

1.2.1. Craniofacial diversity in different human populations

A primary visual characteristic of humans is the anatomy of their facial shape and features. Most individuals are able to recognise the specific differences in facial appearance between various people. In the process of facial recognition, people do not rely solely on facial morphology, but also use additional information available, such as skin, eye and hair colour. While perceiving a face, people often try to distinguish a person's ethnicity as well. In this process, some facial features together with facial pigmentation would become the primary factors in categorising someone's ethnicity and facial recognition. These factors form a complex image, which is recorded in the memory and used when needed [15]. However, the information perceived by human brain is not always adequately remembered. Some details may be forgotten or

remembered wrongly. In the forensic context, this may lead to apocryphal information provided by eyewitnesses and point the investigation in the wrong direction [4, 5, 7, 16]. As a result, retrieving scientifically reliable information about the ethnic origin and physical appearance of a person may verify eyewitness description and help with crime investigation.

Physical anthropologists have long been aware of variation in facial measurements among different ethnic groups [17, 18]. However, originally most studies of this topic were non-systematic and represented only by visual observations, which lacked a standardised scientific basis until the early 18^{th} century [17-19]. Starting from approximately 30 years ago, these measurements have been systematically studied in various ethnic groups, which had not previously been compared [20-22].

In addition to ethnic variation, there is an obvious sexual dimorphism in the craniofacial morphology. In general, males have larger facial features than females, although the reason for this dimorphism is not clear [17, 20, 21]. The most likely explanation is that the facial morphology is correlated with sexual dimorphism in body size and oxygen consumption. In fact, a recent study demonstrated that the larger noses in males significantly correlate with greater body mass and are needed to increase the oxygen intake and consumption [23].

The size and the shape of facial features are known to be under evolutionary pressure and some of them may even predict human behaviour or medical conditions. In addition to the most obvious influence of overall facial attractiveness on the mating choice and other social interactions, particular facial features may have significant psychological impact on humans' social behaviour. Intuitively, this may include the shape and size of the chin (cleft or "strong"), straight or concave nose and thin or thick lips. Haselhuhn et. al. found that men with wider faces (specifically with greater width-to-height ratio) are perceived by others as more likely to lie, being aggressive and make other people act selfishly [24]. Another study showed that brown-eyes men are perceived as more dominant than blue-eyes men in the European population data set [25]. Interestingly, males with brown eyes have statistically significant broader and massive chins, broader mouths, larger noses, and eyes that are closer together with larger eyebrows. Other studies found association between head and face form (specifically increased cephalic index and decreased facial index) and brachycephaly with increased risk of obstructive sleep apnea or apnea hypopnea index respectively in the Caucasian population [26, 27].

Predicting several aspects of human behaviour, known to be in correlation with craniofacial features and derived from a DNA sample may provide additional useful information for crime investigation, although the genetics behind this link is still unclear. Since many human craniofacial anthropometric characteristics display a normal distribution (within a single healthy population), mean values appear to be useful in describing these characteristics [17, 18]. Figures 1 and 2 show average facial images of men and women of various nationalities, which were produced by merging approximately 50 images for each ethnicity, as a part of the "Face of Tomorrow" project (www.faceoftomorrow.com). Comparing the size and shape of facial features between the images, may provide useful information on the specific differences between these nationalities. While the differences in facial appearance between various ethnicities are easily recognisable, they only represent average features, which need to be "translated" into accurate anthropometric measurements to provide a useful tool in forensic investigation.

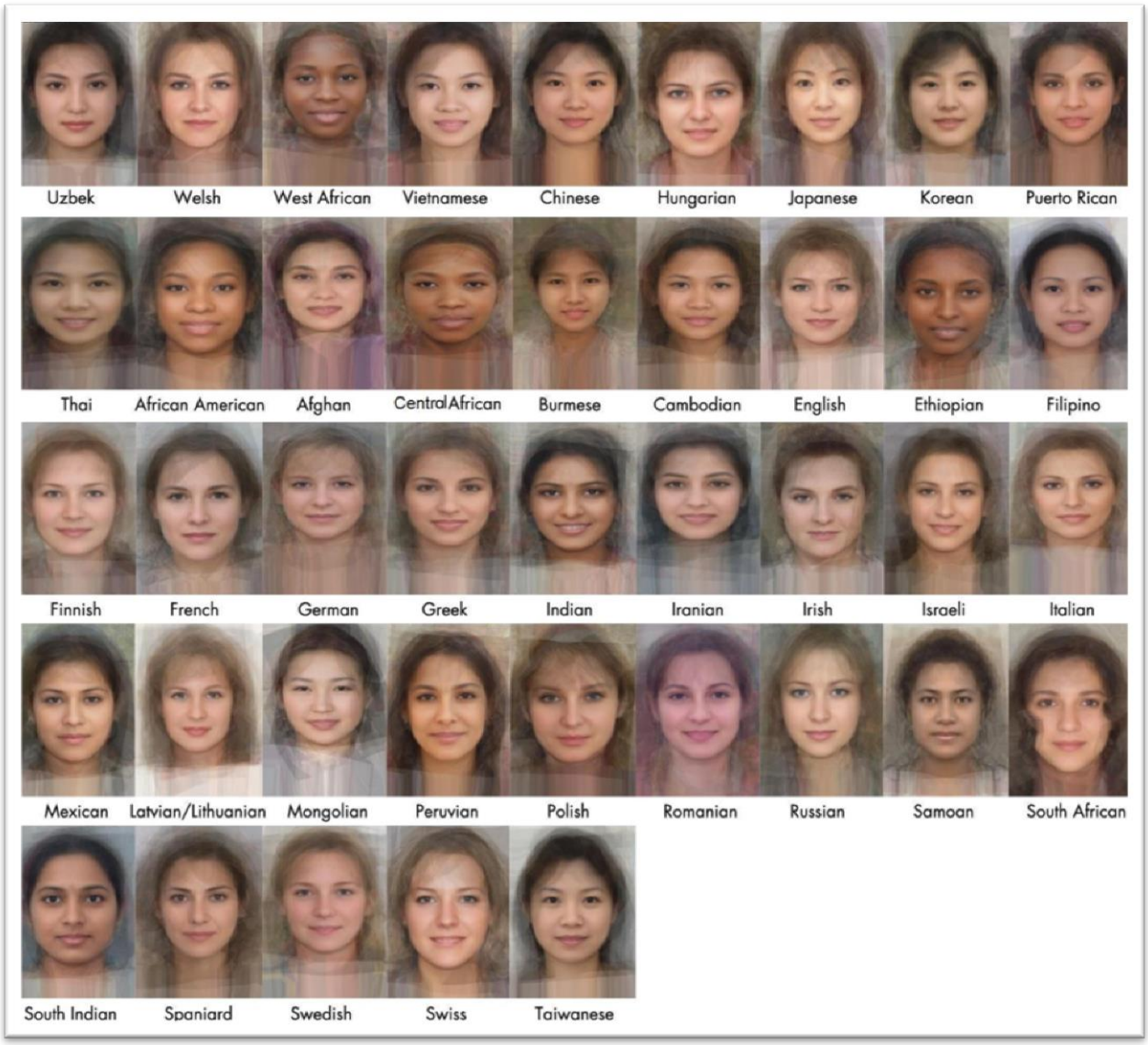


Figure 1. Average female face in different nationalities.

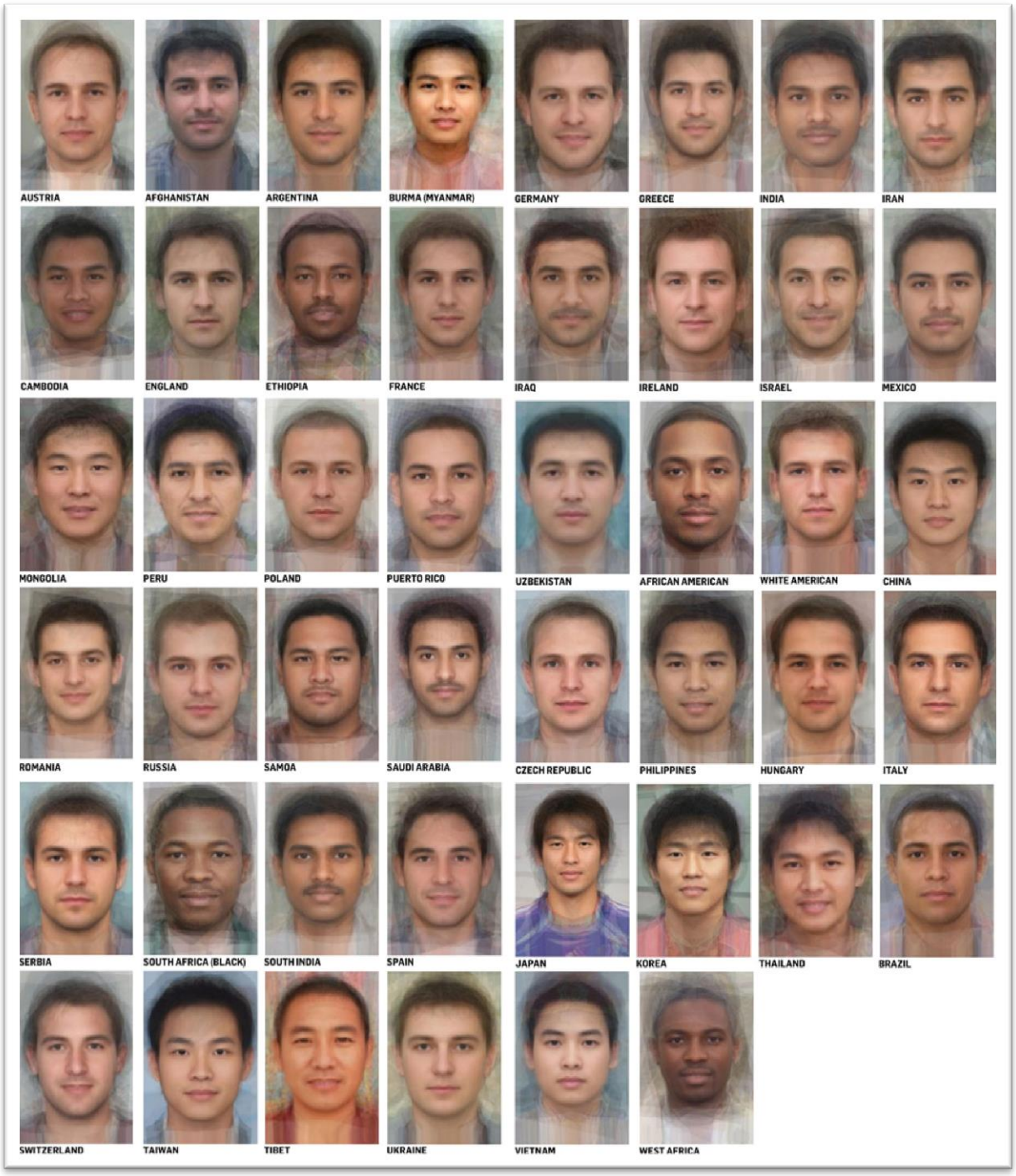


Figure 2. Average male face in different nationalities.

1.2.2. Anthropometric measurements of the head and face

The following section briefly describes basic craniofacial landmarks on the skull and the soft facial tissue as well as various measurements derived from them. All specific measurements relevant to this work are described in details in the Chapter 2.

1.2.2.1. Cranial (hard tissue) landmarks and measurements

The cranium is a support matrix for the facial soft tissue, although its morphology is not a direct indication of the overlying facial structures. The bone is covered by a soft tissue layer, making it difficult to find most of the cranial landmarks on a living subject.

There are a number of fundamental cranial landmarks, providing descriptive information about the cranium such as: Bregma (b), Euryon (eu), Gnathion (gn), Gonion (go), Nasion (n), Opisthocranion (op), Vertex (v) and Zygion (zy), as detailed in Figures 3-5. These landmarks form the basis for numerous linear and angular measurements as well as over 160 indices (ratios between these measurements), which provide information on facial proportions, established largely by Farkas [17, 18]. There are more than 150 facial proportions described in the literature [18]. Not surprisingly, some proportions are significantly different between ethnicities and play an important role in describing facial aesthetics. Knowing aesthetical proportions in the relevant population is very useful in a case of surgical craniofacial intervention [18, 20-22, 28].

Identification of some of the craniofacial landmarks can be challenging, especially landmarks whose location depends on the position of the head or are described only generally (such as the highest or the most distant point of the head) or are placed on bony prominences underlying the skin (like most soft tissue landmarks). This situation may lead to inter or intra examiner inconsistency in measurements.

Figures 3 - 5 illustrate major cranial landmarks used in craniofacial anthropometry. A more comprehensive definition for each landmark is detailed in the List of Terms and Abbreviation and Materials and Methods Chapter. Most of these landmarks are identified by using a spreading calliper (as a maximum distance between two landmarks) or by palpating.

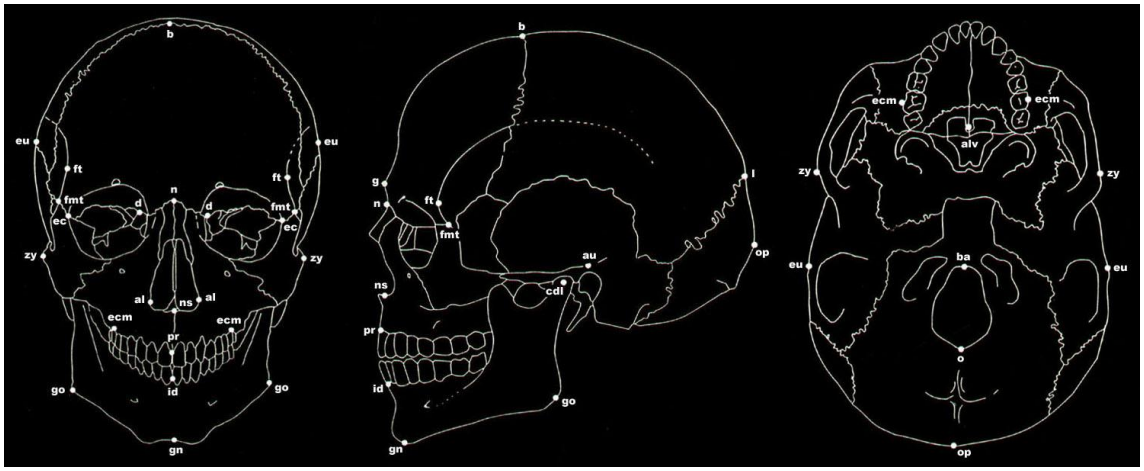


Figure 3. Frontal, lateral and bottom view of skull with cranial landmarks. Sourced from http://www.theapricity.com/snpa/index2.htm.

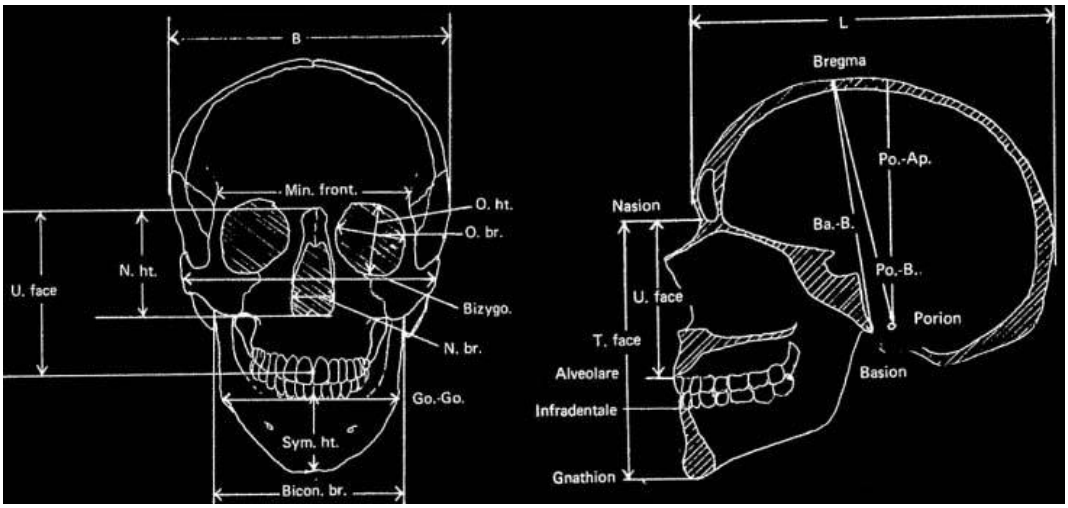


Figure 4. Frontal and lateral view of a skull with major cranial distances. Sourced from http://www.theapricity.com/snpa/index2.htm.

ac	alar curvature point	go	gonion	po	porion
al	alare	id	infradentale	pr	prosthion
au	auriculare	li	labiale inferius	pra	preaurale
b	bregma	l	lambda	prn	pronasale
c'	columella apex	ls	labiale superius	ps	palpebrale superius
cdl	condylion laterale	ls'	labiale superius lateralis	s	sellion
ch	cheilion	m'	nasal midline	sa	superaurale
cph	crista philtre	mf	maxillofrontale	sba	subaurale
d	dacryon	n	nasion	sbal	subalare
ec	ectoconchion	ns	nasospinale	sci	superciliare
ecm	ectomolare	obi	otobasion inferius	sl	sublabiale
en	endocanthion	obs	otobasion superius	sn	subnasale
eu	euryon	on	ophyron	sto	stomion
ex	exocanthion	op	opisthocranion	t	tragion
fmt	frontomalare temporale	or	orbitale	tr	trichion
ft	frontotemporale	os	orbitale superius	v	vertex
fz	frontozygomaticus	pa	postaurale	zy	zygion
g	glabella	pg	pogonion		
gn	gnathion	pi	palpebrale inferius		

Figure 5. Major craniofacial landmarks. Sourced from http://www.theapricity.com/snpa/index2.htm.

Measuring the distance between various landmarks provides a set of craniofacial measurements. The most informative and widely used cranial measurements are:

- **Maximum head length (g-op**)
- **Maximum head height (v-gn)**
- **Maximum head width (eu-eu)**

The location of landmarks, used to calculate these measurements is illustrated in the following picture (Figure 6)

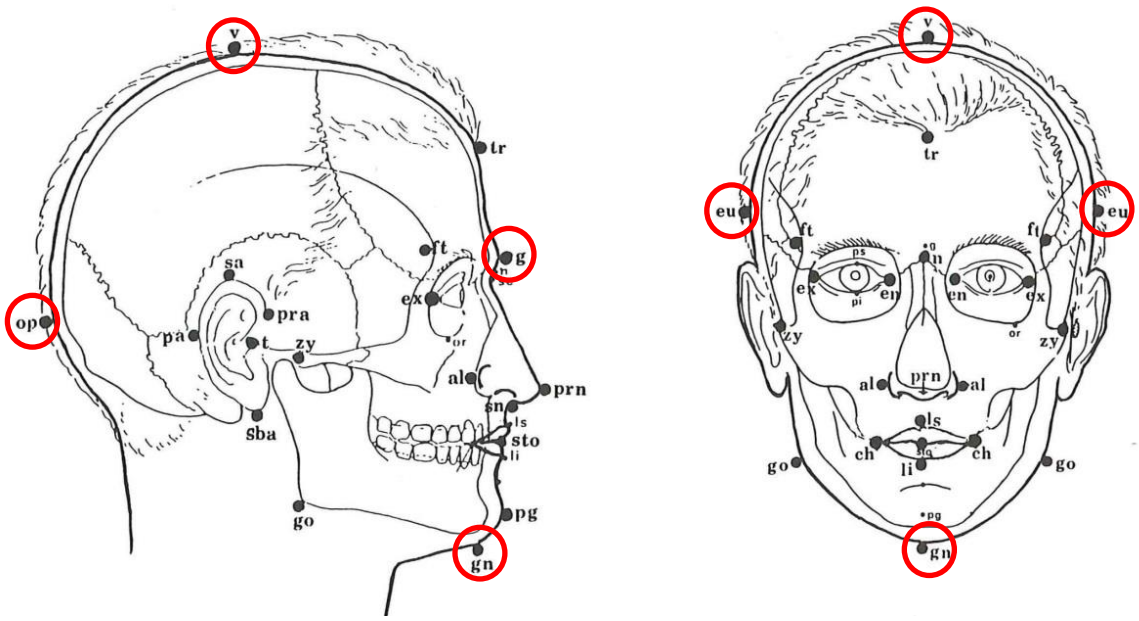


Figure 6. Craniofacial surface landmarks of the head and face in frontal and lateral positions. Figure modified from Farkas L., Anthropometry of the Head and Face (1994) [17].

Two of these measurements are used to calculate the Cephalic index (C.I.), which is considered the most informative craniofacial index. C.I. is defined as: *(Maximum head breadth (eu-eu) / Maximum head length (g-op))* 100.*

The C.I. provides categorical information about head shape and size. According to the Frankfort Craniometric Agreement of 1882 [19], C.I. variation can be categorised as following:

C.I. 55.0 - 59.9 = *ultradolichocephalic* (extremely long-/narrow-headed)
C.I. 60.0 - 64.9 = *hyperdolichocephalic* (very long-/narrow-headed)
C.I. 65 - 74.9 = *dolichocephalic* (long-/narrow-headed)
C.I. 75.0 - 79.9 = *mesocephalic* (intermediate in head form)
C.I. 80.0 - 84.9 = *brachycephalic* (round-/short-/broad-headed)
C.I. 85.0 - 89.9 = *hyperbrachycephalic* (very round-/short-/broad-headed)
C.I. 90.0 - 94.9 = *ultrabrachycephalic* (extremely round-/short-/broad-headed)

The French system [17, 18] however, recognises less categories, omitting marginal values:

C.I. - 75.00 = *dolichocephalic*
C.I. 75.01 - 77.77 = *subdolichocephalic*
C.I. 77.78 - 80.00 = *mesocephalic*
C.I. 80.01 - 83.33 = *subbrachycephalic*
C.I. 83.34 - = *brachycephalic*

Each head type (according to specific CI value), provide additional information on the specific facial morphological traits [1, 17, 29]. A long and narrow dolichocephalic head would most likely demonstrate a protrusive nasomaxillary component to allow more efficient air flow. The nasomaxillary component would be also lowered relative to the mandible, resulting in posterior location of the latter. On the contrary, a wide and short brachycephalic head would most likely demonstrate a short, wide and more posterior located nasomaxillary component, allowing sufficient air flow. The mandible in this case would be located more anterior, protruding from the face. Some head types are more common in the specific population groups than other. Such as dolichocephalic heads are more common in the North African, North European and Middle Eastern populations, while the Asian population is more associated with the brachycephalic heads.

Historically, the European population has been the most extensively studied by anthropologists, with the number of anthropometric studies comparing other population groups being quite limited [22, 28, 30]. In recent years, several studies have compared

additional population groups, such as African and Asian populations [17, 30-34]. These studies demonstrated that based on the C.I. measurements, the Chinese for example, generally have a short head, Europeans generally have a medial head and Africans have a relatively elongated head. Specifically, the greatest difference between these three populations is found in the head height and width. The head in the African population is the largest in length and smallest in width. In the Chinese population, the head is the largest in width and smallest in length. The head in the Caucasian population is generally wider than African, but narrower than in the Chinese population.

Specifically, the mean values of the C.I. may range from the upper limits of dolichocephaly in the Anglo-Saxon, Scandinavian and African populations, through mesocephaly in the Slavic and Central European populations to hyperbrachycephality in the central Asian populations. In spite of the wide range of C.I. variation, the "normal" values range in the world – wide population is relatively limited and usually represented as between approximately 73-75 and 88-89 [28]. Despite the statistically significant difference in C.I. and other craniofacial measurements between various population groups, these measurements are not highly descriptive and prone to potential error in measurements. As a result, they cannot be efficiently used for descriptive population differentiation, especially in the modern cosmopolitan society with a high degree of admixture. This problem can be potentially solved by applying a set of DNA markers, which show strong association with specific anthropometric measurements and bio-geographic ancestry.

1.2.2.2. Facial (soft tissue) landmarks and measurements

Traditional anthropometry treats the face as a two-dimensional object. Most facial measurements cover the vertical plane (facial length), some cover the horizontal plane (facial breadth) and only a few cover the facial depth. The facial depth measurements are the most difficult to capture. However, the use of a 3-Dimentional scanner can provide extensive information of a facial surface and potential measurements (as detailed in Section 1.2.3)

Some of the soft facial tissue landmarks represent the same landmarks, which can be found on the cranium (e.g. glabella, zygion and gonion), while others can be allocated only on a living subject. Figure 7 illustrate classical facial landmarks based on Farkas et.al [17].

a) b) c)

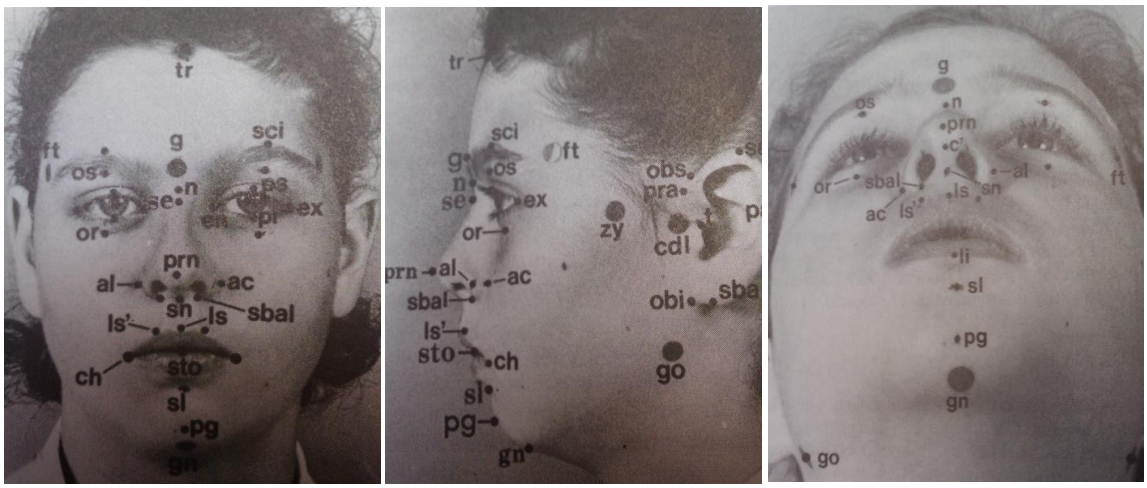


Figure 7. Craniofacial landmarks of the soft tissue on a- frontal, b – lateral, c- base aspects. Sourced from Farkas et.al. [17].

In the traditional anthropometry, the direct assignment of the craniofacial landmarks is an integral part of the measurements. There is a need for location of the craniofacial landmarks by palpating, followed by manual measurement of the linear and angular distances between these landmarks using a calliper. In the 2D or 3D facial images however, the location of landmarks and relevant measurements performed indirectly. A difference between these approaches may introduce errors in the digital approach as well as significant challenges if a comparison between the measurements obtained by different methods is required.

Figures 8 and 9 illustrate basic linear measurements of the head and face. A detailed list of the craniofacial landmarks used in this study is presented in the Chapter 2.

Facial measurements of the vertical plane:

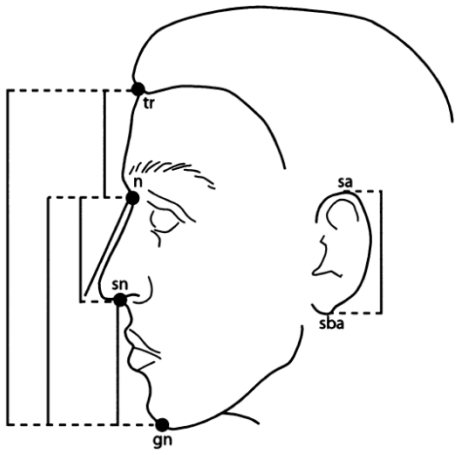


Figure 8. Perpendicular measurements of the face. Sourced from Farkas et.al. [1].

Physiognomic facial height (tr-gn)

Forehead height (tr-n)

Nose height (n-sn)

Morphological face height (n-gn) (total face height)

Lower face height (sn-gn)

Ear length (sa-sba)

Facial measurements of the horizontal plane:

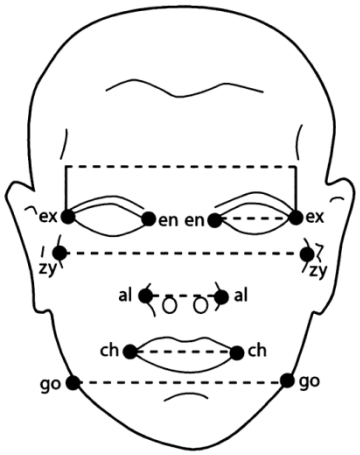


Figure 9. Horizontal measurements of the face. Sourced from Farkas et.al. [1].

Binocular width (ex-ex)

Maximum facial breadth (zy-zy)

Nose width (al-al)

Mouth width (ch-ch)

Mandible width (go-go)

Linear measurements, illustrated in Figures 8 and 9 can be used for calculating several indexes. Among others, the facial index (F.I.) is an important basic facial measurement, also known as total facial index, which is calculated as follows**:** morphological face height (n-gn) / maximum facial breadth (zy-zy)* 100.

According to Farkas [18], facial index measurements can be divided into four main groups, which used to categorize human faces:

F.I. - 79.9 = *hypereuryprosopic* (very short-/broad-faced)
F.I. 80.0 - 89.9 = *euryprosopic* (short-/broad-faced)
F.I. 90.0 - 94.9 = *leptoprosopic* (long-/narrow-faced)
F.I. 95.0 - = *hyperleptoprosopic* (very long-/narrow-faced)

The nose is one of the most prominent facial features. Nose width and height show significant variation between various ethnicities. The ratio between nose width and height is represented by Nasal Index (N.I.), which is calculated as follows: nose width (al-al)*100 / nose height (n-sn).

According to the Frankfort Agreement [19] N.I. can be categorized into four main groups:

N.I. - 47.0 = *leptorrhine* (narrow-nosed)
N.I. 47.1 - 51.0 = mesorrhine (having a nose of moderathe width)
N.I. 51.1 - 58.0 = platyrrhine (wide-nosed)
N.I. 58.1 - = hyperplatyrrhine (very wide-nosed)

A comparison between the mean values of various population groups show that the Chinese population (for example), generally has a more euryprosopic face with a platyrrhine nose, while the European population has a more leptoprosopic face with a mesorrhine or leptorrhine nose and the African population tend to have a more euryprosopic face and a platyrrhine to hyperplatyrrhine nose [21, 22, 28]. Specifically, European noses tend to be longer and narrower than African and set more highly on the face. Europeans usually have narrow faces, higher cheekbones, wider cranium and more prominent foreheads than Africans. In addition, jaws in the African population protrude more from the face than do European jaws. Africans also have wider orbits and wider inter-orbital spaces than Europeans. Asians typically have wide and flat faces, wider spaced eyes and flat supranasal regions. In general, Europeans have smaller faces than most non-European populations, in spite of being taller and heavier than many non-European populations.

1.2.3. Methods of facial landmarks analysis

There are two types of craniofacial anthropometrical measurements: direct and indirect. Direct anthropometric measurements involve using traditional measuring equipment such as sliding and spreading callipers and measuring tape. The advantage of direct measurements is that this process is objective, reliable and inexpensive. The examiner has full access to all the bony landmarks, reducing potential errors in incorrect location of these landmarks in indirect methods. The main limitation of direct measurements is that the examination may take approximately an hour or even longer, making it ineffective for large sample size and working with children.

The indirect approach involves using 2D or 3D images for performing craniofacial measurements. Since the publication of the first English-language manuscript on anthropometry by Ales Hrdlička in 1920 [35], which introduced a set of 14 cranial and facial measurements, the methods for accurate measurements of the craniofacial features have undergone significant changes. From using manual and time consuming devices such as a calliper and measuring tape, to more sophisticated, accurate and faster methods, such as 3-Dimentional image-capturing methods, represented by Computer Tomography (CT), Magnetic Resonance Imaging (MRI) and more recently by a 3-Dimentional (3-D) laser scanning [36-43].

The advantage of the indirect measurement is that it takes significantly less time than the direct approach (few minutes) and offers better resolution (such as CT and MRI), providing an opportunity to collect more samples and also use this technique in the medical field [37, 44]. Using regular photographs (2D images) for collecting indirect measurements is also possible. This method is fast, cheap and easy. However, 3D image capturing technologies provide a more accurate image than 2D photos, with a computer file output, that can be processed further by automatically calculating craniofacial measurements. In addition, 2D photographs do not reflect the natural anatomy of the face and as a result do not provide information on the facial surface, omitting useful information. The CT scan offers a relatively high image quality, but involves exposure to low levels of radiation. This technology is also expensive and not easily accessible. MRI provides a very high resolution soft tissue representation without exposing patients to radiation, but it is even more costly than CT, is not easily accessible and is also time consuming. 3D laser-scanning on the contrary, is a completely health-safe, fast, easy and cost – effective procedure, offering high-resolution facial images. 3D laser digitizers have been validated in several studies, showing very high precision and have

been demonstrated to accurately capture even minor facial appearance differences in twins [38, 41, 45-47].

As an alternative to the traditional direct measurements, the use of digital equipment allows extraction of the Euclidean coordinates (x, y and z) for each landmark and subsequent calculation of various distances in an automatic and fast manner, without the need to perform each measurement separately.

Based on the Euclidean geometry rules, the distance between two coordinates in the 3D space can be calculated using the following formula:

Linear distance = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$

Given that the distance between points 1 and 2 is "a", between points 2 and 3 is "b" and between points 1 and 3 is "c", the angle between points 1, 2 and 3 can be calculated by using this formula:

Angular distance = $\frac{cos^{-1}(a^2+b^2-c^2)}{2(a*b)}$

These formulae can be programmed in to user-friendly software, such as an Excel spreadsheet, and used for relatively easy and fast extraction of various craniofacial measurements based on landmarks coordinates.

However, digital technologies have a few limitations. The main limitation of the 3D (as well as 2D) image-capturing techniques is the image processing, which includes facial landmark location and extraction. Performing this procedure manually is tedious and time-consuming. Automatic facial image processing has been explored in several studies, but is still under development and standardization [36, 37, 44, 47-49]. It involves application of special algorithms, which recognize specific anatomical features and craniofacial landmarks. This approach usually involves using a set of images that have been processed manually for location of these landmarks, which are used by a computer program to automatically recognise the same landmarks on a new image based on the similarity. However, current algorithms used for this purpose are still not able to accurately recognize facial landmarks on various faces. Given that each face is different from another, this approach does not always produce an accurate outcome. As a result, most landmarks generated using this approach have to be re-examined manually, which may significantly decrease the advantage of this method versus manual

landmark allocation. In addition, most of the algorithms used for this purpose are "in-house" developments and have not been extensively validated or are easily available for use. The majority of these algorithms are available as "command-line" software, requiring additional programming skills knowledge.

1.2.4. Craniofacial reconstruction

Facial reconstruction, which uses anthropometric measurements from skull remains to estimate the facial appearance of a person is an important tool in forensic investigation. Facial reconstruction is based on the fact that the bony cranium represents the supporting basis for soft facial tissue [50]. However, there is no precise correlation between the cranium and the soft tissue. The reconstruction of the nose, mouth and ears are particularly difficult, as these soft tissues are not supported by a bony frame and the cartilage under these facial features is often not present in skeletal remains [51, 52]. In addition, other phenotypic features such as pigmentation and hair texture cannot be reconstituted from anthropometric measurements and are usually a best guess by the anthropologist, based on inferred ancestry. As a result, most current facial reconstruction methods focus on generating an average face, which may or may not resemble the original one [52-54]. With the recent development of more accessible 3-D image – capturing technology this process has become more accurate, although it is still far from being standardised [46, 51, 55, 56].

In some forensic cases the skull is not available for analysis. In such cases, the DNA analysis is essential. While DNA typing of skeletal remains is routinely used for identification or paternity purposes, the recent advances in human pigmentation genetics, offer an opportunity to predict of eye, skin and hair colour from skeletal remains [9]. A better understanding of the craniofacial genetics and in particular, knowing the genetic factors behind the size and shape of the specific craniofacial features, would allow potential prediction of these traits and more accurate craniofacial reconstruction from a skull or even a little piece of biological (even non-skeletal) evidence.

1.3. Craniofacial development and embryogenetics

This section briefly outlines the major stages of the craniofacial embryonic development in *Homo sapiens*, the genes which regulate this complex process and discusses craniofacial malformations that may occur as a result of genetic mutation in these genes.

Human craniofacial development is a complex multistep process, involving numerous signalling cascades of factors that control neural crest development, followed by a number of epithelial-mesenchymal interactions that control outgrowth, patterning and skeletal differentiation. The mechanisms involved in this process include various gene expression and protein translation patterns, which regulate cell migration and positioning. These events are precisely timed and are under hormonal and metabolic control. Most facial features of the developing human embryo are recognizable from as early as 5 weeks post conception (Figure 10). The resulting face is different from person to person, even in the case of monozygotic twins, whose faces (the soft tissue) acquire differences with the age – most likely, as a result of the epigenetic influence on phenotype. This influence may be a function of various factors, including differences in nutrition (influencing hormones and growth factors) as well as social interaction (e.g. various muscles involved in facial mimics) [57-60].

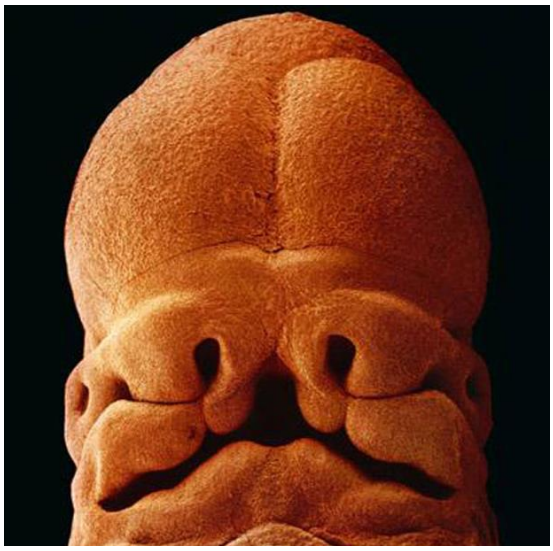


Figure 10. Human embryo at 5 weeks showing early stages of facial feature formation. Eyes, nose and mouth are easily recognizable. Courtesy of Lennart Nilsson [61].

A comprehensive overview of different stages of foetal development made by *in vivo* video animations is available from Lennart Nilsson's web site: http://web.tt.se/lennart_nilsson_video/index_m.html.

The vertebrate head is a composite structure whose formation begins early in development, as the brain begins to form (Figure 11).

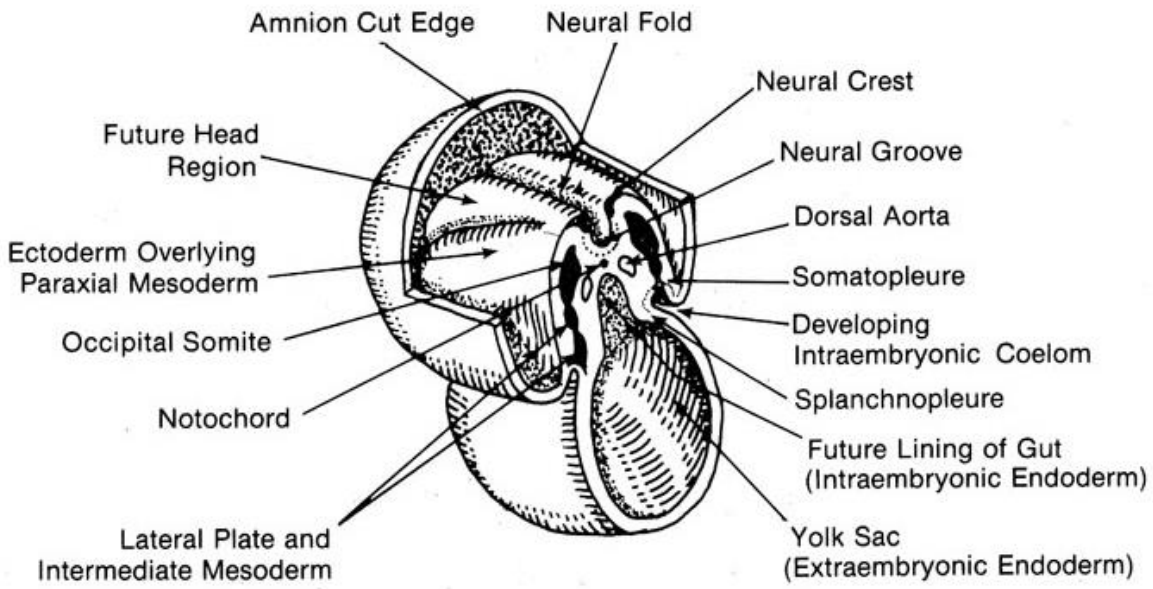


Figure 11. Transverse section through 20-day-old embryo depicting neural folds and neural crest formation. Sourced from Sperber et.al.[62] .

Central to the development of the head is the concept of segmentation, evident in development of the hindbrain and branchial arch systems. Together with migrating neural crest cells, these systems give rise to most of the head and neck compartments (Figure 12).

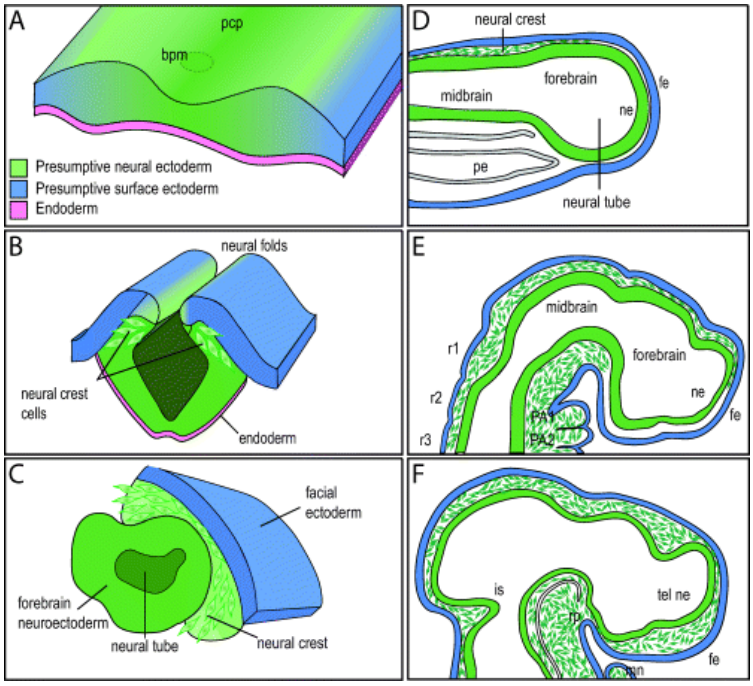


Figure 12. Neural crest induction and migration in the developing embryo. (A) The neural plate consists of a unified layer of ectoderm, beneath which lies the endoderm. The neural folds arise as the ectoderm begins to fold upwards. Interactions between signalling molecules cause the medial portion of ectoderm to begin to assume a neural character (green) while lateral portions of ectoderm begin to take on a non-neural character (blue). The prechordal plate mesendoderm (pcp) and the buccopharyngeal membrane (bpm) are indicated. (B) As the neural folds begin to fuse, the neural tube takes shape, giving rise to distinct tissue layers of neuroectoderm (green) and surface ectoderm (blue). Neural crest cells start to delaminate from the border region between the neuroectoderm and surface ectoderm. (C) Once the neural tube has closed, neural crest cells lie interposed between the facial (surface) ectoderm (fe) and the neuroectoderm (ne). (D–F) As the central nervous system begins to form from the neural tube, the neural crest starts to migrate anteriorly from rhombomeres (r1–r3) into different areas of the face, and into the pharyngeal arches. Abbreviations: C, caudal; is, isthmus; mes, mesencephalon; mn, mandible; PA, pharyngeal arch; pe, pharyngeal endoderm; rp, Rathke's pouch; R, rostral; tel ne, telencephalic neuroectoderm. Based on Tapadia et.al. and Helms et. al. [63, 64].

The neural crest is a pluripotent cell population that plays a critical role in the development of the vertebrate head. This ectomesenchymal tissue arises from the crests of the neural fold during the gastrulation process of the embryonic disk (Figure 11). Unlike most parts of the body, the facial mesenchyme (whole viscerocranium and part of the neurocranium) is formed principally from the neural crest and not the mesoderm of the embryonic third germ layer [63, 64]. The cranial part begins its development as early as the middle of the third week post coitum. By the fourth week, neural crest cells migrate extensively throughout the embryo in four overlapping domains: cephalic, trunk, sacral and cardiac. Subsequently, the cephalic neural crest cells migrate from the posterior midbrain and hindbrain region into the branchial arch system. The ectomesenchymal neural crest cells then interact with epithelial and mesodermal cell populations present within the arches, leading to the formation of craniofacial bones,

cartilage and connective tissues. The dermatocranium (roof of the skull), is however formed not from cartilage, but from direct ossification of the deep layers of the dermis. Cells, which migrate within the cranial paraxial mesoderm form somitomeres, which subsequently develop into the muscles of the face and jaws. Other neural crest cell populations provide mesenchyme for angiogenesis to produce blood vessels, in addition to others, which will develop later to melanocytes for skin and eye pigmentation.

The initial orofacial development is induced by two organizing centres. The first is called prosencephalic centre and is derived from prechordal mesoderm. It induces the development of eyes, inner ear and upper third of the face. The second is called rhombencephalic centre, and induces the middle and lower thirds of the face (Figures 13 and 14). Simultaneously with these processes develops the forebrain, which induces multiple signalling areas in the ectoderm. These areas control the differential cell proliferation of the nasal area in the upper face region. Subsequently, these signalling cascades induce the ectomesenchyme to develop five prominences, such as paired maxillary and mandibular and single frontonasal, which give rise to the specific facial features.

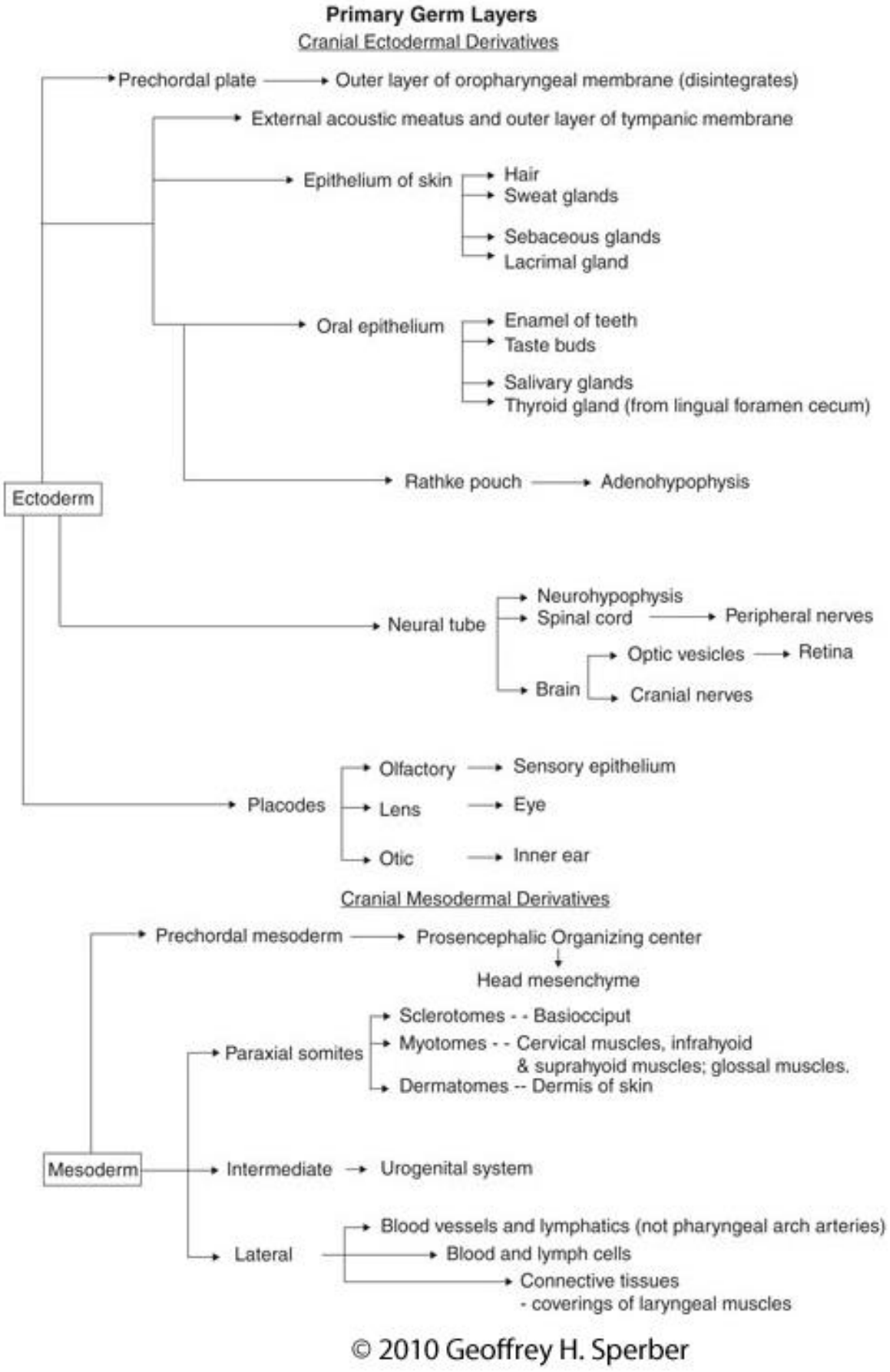


Figure 13. Primary germ layers and their derivatives. Sourced from Sperber et.al., [62].

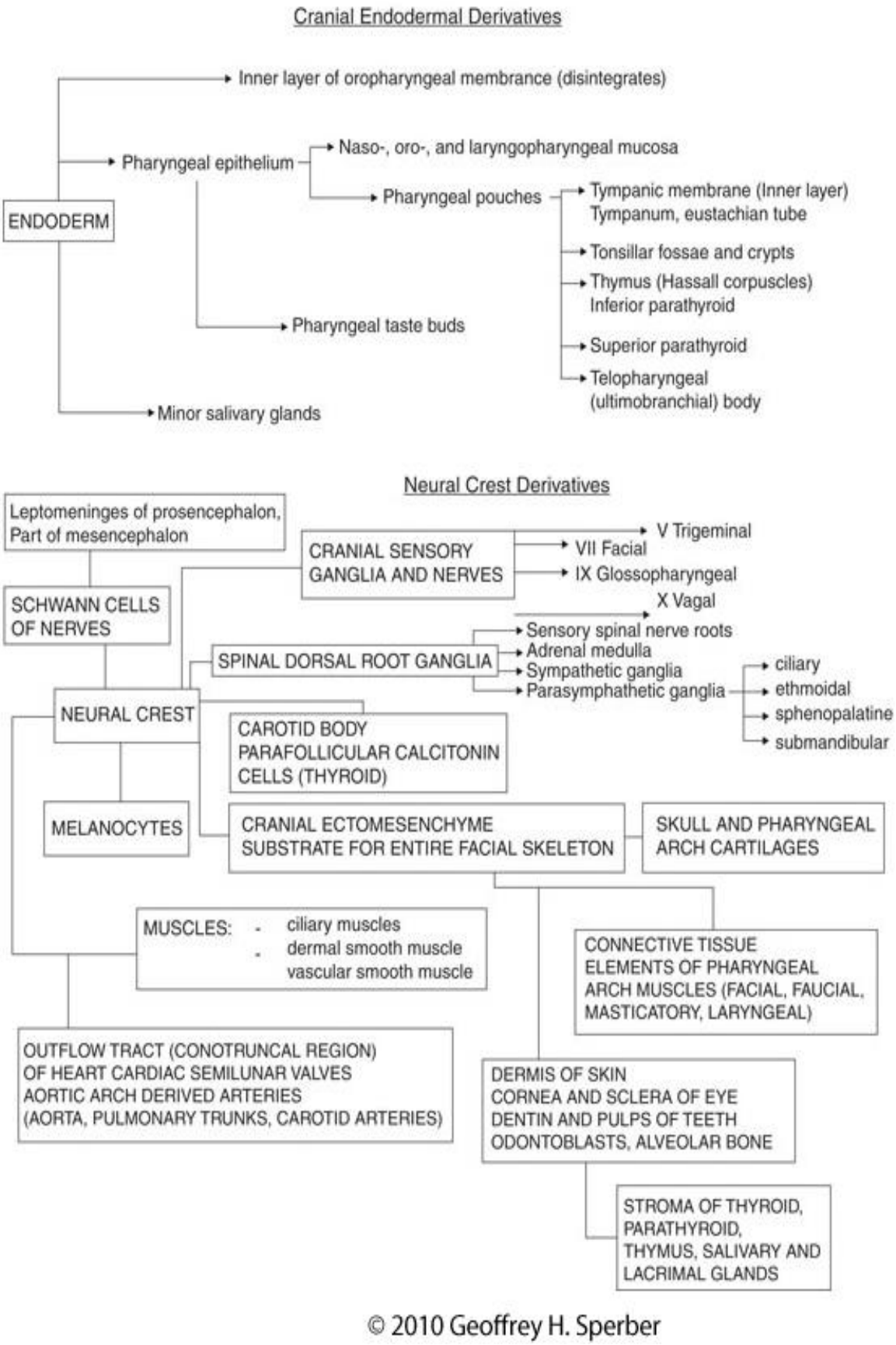


Figure 14. Primary germ layers and their derivatives. Sourced from Sperber et.al., [62].

The developing head is a community of all of these cell populations (Figure 14). Disruption to any one of these pathways, interfering with a normal craniofacial three-dimensional developmental process, may lead to a variety of craniofacial syndromes including disruption of brain morphogenesis, usually accompanied by facial malformations.

Interestingly, there are a number of known malformations of the limbs, which are expressed together with various craniofacial abnormalities [65-68]. This is most likely due to the overlap in signalling pathways involving limb and head development. Therefore, genes that are involved in limb development may also influence normal variation in craniofacial features.

1.3.1. Craniofacial embryogenetics

The cranial and facial tissues are comprised of a large number of complex structures whose development is controlled by a wide number of genes, expressed in a specific pattern during embryonic maturation (Figure 6). These genes represent different gene classes (many of which can be differentiated into further subcategories) and include:

- **Homeobox domain genes** such as: distal-less homeobox (DLX), hedgehog (HH), wingless (WNT) and aristaless-like homeobox (ALX) gene families;
- **Signalling and growth factors and their receptors** such as: fibroblast growth factor and its receptor (FGF and FGFR), epidermal growth factor and its receptor (EGF and EGFR), transforming growth factor β (TGFβ), endothelin 1 (ET1), Bone morphogenetic protein (Bmp) antagonists (e.g. Chordin and Noggin) and Jagged;
- **Transcription factors** such as: ALX, BARX, BMP, FGF, DLX, GSC, LHX, MHOX, MSX, PAX, PITX, TP63, gene families;
- **Matrix proteins** such as: Collagen, ANKH, FGD1;

These genes can be also classified by their expression in various stages of embryonic development of particular facial tissues (Figure 15). For example, the genes ALX4, BMP2,4,7; BSP, COL1, DLX5, FGF1-3, FGFR1-3, FOXC1, ID1, MSX1, MSX2, PTC, RUNX2, TWIST, TGFβ1-2 and SHH are involved in cranial sutures development; ALK5, BMP, FGF, Lrp6, SHH, TGFβ1, WNT are involved in nasal pit, upper lip and upper jaw formation and DLX and PAX6 are involved in eye formation.

The majority of these genes have been identified in animal models, expressing various types of craniofacial malformations. Although many of mutations in these genes can lead to facial anomalies, currently there is little information available regarding specific polymorphisms, which can be identified as potentially indicative of normal variation in craniofacial shape and other facial characteristics.

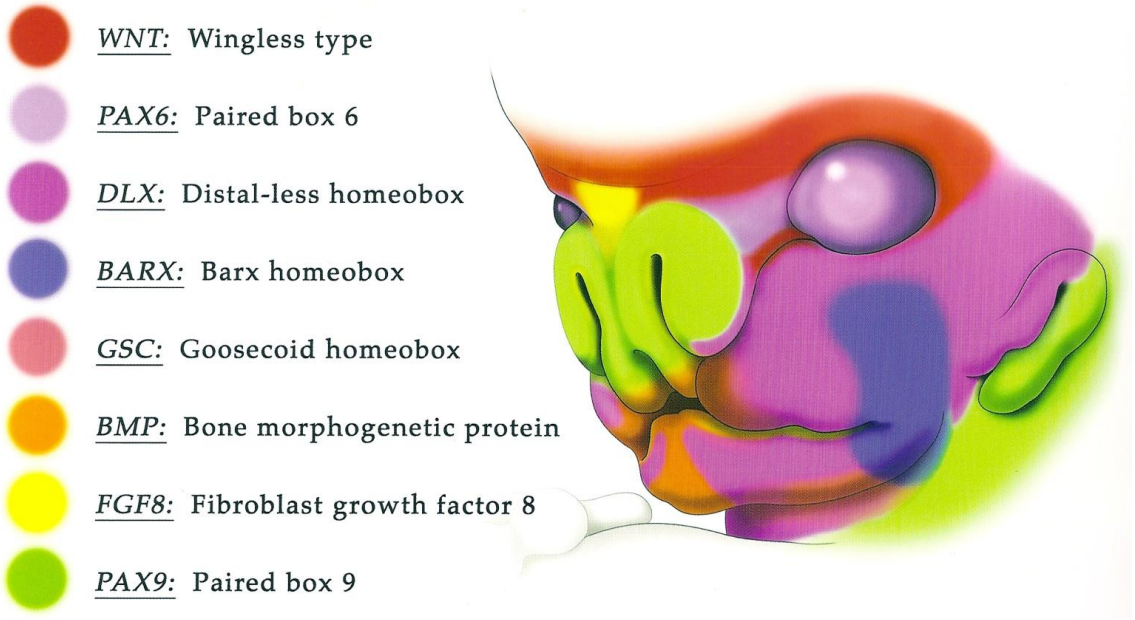


Figure 15. Keyed drawing of gene expression patterns of developing embryonic face of approximately 7 weeks post conception age. Sourced from Sperber et.al., [62].

1.4. Possible approaches for finding genes and markers associated with normal variation in the craniofacial appearance

The identification of the markers responsible for normal craniofacial appearance can be achieved by taking two main approaches:

1. Genome Wide Association Study (GWAS).

This approach usually involves genotyping millions of SNPs in thousands of individuals and analysing the data for potential associations between the trait of interest and genetic markers [69]. This approach has successfully identified genes and mutations involved in

many complex disease and traits [70, 71]. In spite of being a very powerful approach, the requirement for such a large sample size and the need to type millions of genetic markers, as well as the high cost, renders the approach not feasible for this research project.

2. Candidate gene approach.

This method does not require as large a sample size as GWAS approach, due to a more focused genotyping of a limited number of markers. It can be performed using one or all of the following pathways:

- Identification of genes previously shown to be involved in normal human facial appearance and subsequently selecting markers, which may have an effect on the phenotype (as detailed in section 1.4.1).
- Genotyping markers in genes, shown to be involved in craniofacial disorders in human and model organisms. Genes that result in disease phenotype may also affect normal variation of that phenotype, similar to genes involved in Albinism and normal pigmentation [72] .
- Genotyping markers showing high population differentiation (high Fst values) and particularly in or near genes, potentially involved in craniofacial embryogenesis. Based on visual differences in craniofacial appearance between various populations, it may be hypothesized that markers with high Fst values, which are located in genes expressed in various stages of craniofacial embryonic development, would be involved in determination of the normal craniofacial appearance.

Based on the cost and time limitations of this project, a candidate gene approach was chosen as the most appropriate method for finding the SNPs involved in the craniofacial appearance.

1.5. Genetics of the normal craniofacial development in human

Until recently, molecular cascades that control human craniofacial development, and particularly of specific facial features, have been poorly understood. In fact, prior to 2010 there were only three studies that described associations between specific

craniofacial measurements and SNPs in only few genes such as GHR, FGFR1 and ENPP1 genes [73], [74], [75]. In 2012, a further two studies identified an association between several markers and specific craniofacial anthropometric measurements [76, 77]. A GWAS undertaken in the same year found that PAX3 gene, previously associated with Waardenburg Syndrome, influences the growth of nasal tissue and specifically the 'nasion' position. SNP rs7559271 allele of PAX3 gene was found to be associated with an increase of 0.39 mm in nasion-mid endocanthion distance [78]. The second study identified associations between various facial measurements and SNPs, such as between rs4648379 in PRDM16 and alare-pronasale distance; between rs974448 in PAX3 and eyeball - nasion distance; between rs17447439 in TP63 and left and right eyeballs distance; between rs6555969 in C5orf50 and zygion – nasion and eyeball – nasion distances; and between rs805722 in COL17A1 genes and eyeballs and nasion distance [77].

The limited number of studies on the normal human craniofacial appearance illustrates the demand for a complimentary approach in searching for the craniofacial candidate genes and markers.

1.5.1. Craniofacial syndromes with known genetic mutations in human and model organisms

The cranium is composed of two major groups of bones: those that surround the brain called neurocranium (composed of desmocranium and chondrocranium) and those that surround the oral cavity, pharynx and respiratory passages known as viscerocranium. Neurocranium gives rise to the bones of the cranial vault, known as calvaria. At birth, the cranial vault surrounding the brain consists of seven calvaria bones with adjoining bones separated by fibrous joints (growth centres) called calvarial sutures (Figure 16). Interestingly, the bones of the jaws and facial skeleton (viscerocranium) are evolutionary younger than chondrocranium and more susceptible for developmental defects [79]. While serious malformations in these structures may lead to different facial abnormalities, polymorphisms in genes responsible for the development of the cranial vault and specifically neurocranium, might influence normal cranial shape [80, 81].

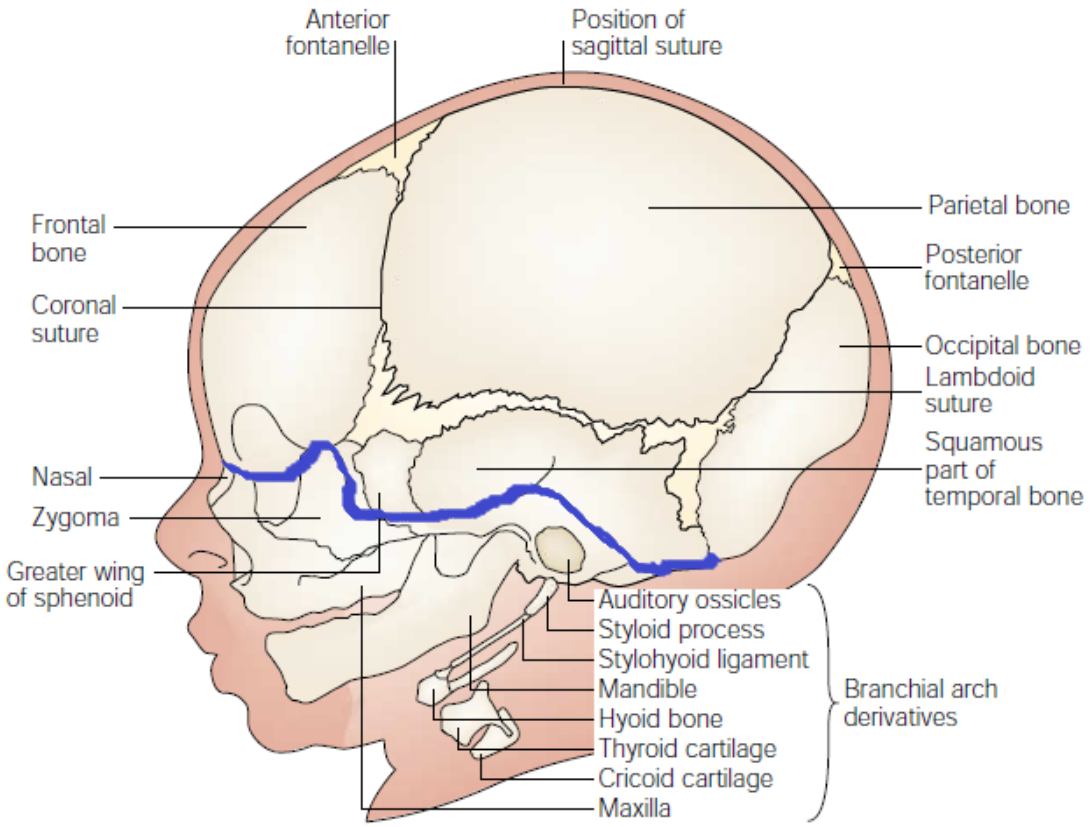


Figure 16. Anatomical features of the human skull at birth .The blue line indicates an approximate border between the neurocranium and viscerocranium. Sourced and edited from Wilkie et.al., [79].

There are at least 250 syndromes, that display various craniofacial abnormalities such as anencephaly, holoprosencephaly, exencephaly, as well as skeletal and facial defects, such as cleft lip/palate, (discussed in more detail in Section 1.4.3). The facial defects may be restricted to specific regions of the head, although are often linked to cardiac and aortic arch defects or developmental limb defects, reflecting the common origin of these structures.

One of the most common craniofacial defects is craniosynostosis, caused by premature fusion of calvaria bones. Craniosynostosis can occur sporadically and termed non-syndromic or be inherited (usually as a dominant trait) and termed syndromic [82, 83]. Syndromic craniosynostosis usually involves additional abnormalities of the skull, as well as normal phenotypes such as long or beaked nose, protrusive or retrusive jaws, wide eyes and large brow ridges. As a result, a research on syndromic craniosynostosis disorders may provide additional insights on the normal craniofacial morphology [74].

Examples of such conditions include DiGeorge syndrome (1 in 4,000 births), Treacher-Collins syndrome (1 in 10,000 births), Saethre-Chotzen syndrome (1 in 25,000 births), Pfeiffer syndrome (1 in 100,000 births) and Rieger syndrome (1 in 200,000 births), all of which are inherited in an autosomal dominant manner [84-96]. Malformations of the limbs and/or the skull have been associated with mutations in several human and model organism genes, such as sonic hedgehog (SHH), fibroblast growth factor (FGF) receptor family and transcription factors like TWIST [62, 79].

A number of other genetic syndromes are known to affect facial connective tissue and specifically the cartilage. These disorders are generally known as Collagenopathy type II and XI and caused by mutations in COL2A1 gene (type II collagen) or in COL11A1, and COL11A2 genes (type IX collagen). Various mutations in collagen genes may result in a number of conditions, such as Hypochondrogenesis, Achondrogenesis type 2, Stickler syndrome, Kniest dysplasia, Weissenbacher–Zweymüller syndrome, Otospondylomegaepiphyseal dysplasia and Spondyloepimetaphyseal dysplasia [97-100].

In addition to "classic" craniofacial disorders, there are many syndromes that are not formally designated as craniofacial, but among other major symptoms, show distinct dysmorphic facial features which are often used to diagnose these disorders. This group includes conditions with various chromosomal aberrations such as Down syndrome, Fragile X Syndrome and Turner Syndrome [101-108] as well as specific mental conditions, such as schizophrenia [109-111]. For example, people suffering from Down syndrome demonstrate brachycephaly, smaller and flatter cranial base, almond-shaped eyes, epicantic folds, reduced orbital height and width, small nasal bones, mandibular prognathism and ear dysmorphology. Facial features of the Fragile X Syndrome are characterised by long and narrow face, prominent forehead and large ears. Females patients with Turner syndrome demonstrate distinct facial features, such as downward slanting eyes, prominent earlobes and very broad necks. Patients suffering from schizophrenia demonstrate narrowing and reduction of the mid to lower face and frontonasal prominences, including reduced width and posterior displacement of the mouth, lips, and chin, increased width of the upper face, mandible, and skull base, with lateral displacement of the cheeks, eyes, and orbits and greater downward displacement of the tip of the nose. Interestingly, the association of 22q chromosomal region with schizophrenia by Scutt et.al. [109], which was subsequently confirmed and extended by Xu et.al. [112], may in fact be an influence of nearby craniofacial genes/SNPs, rather than genes involved in schizophrenia itself, although this hypothesis was not considered

by the researchers. This is rather strange, as 22q11 deletion is primarily associated with Di George syndrome, which among other features characterized by elongated facial features often with flat cheekbones, a long 'strong' nose with a relatively broad and prominent nasal bridge, small nostrils and a small jaw. In fact, approximately 25% of adults with 22q11 deletion have schizophrenia [113].

The great number of craniofacial disorders may provide an indication of the number of genes involved in craniofacial embryogenesis and may partially explain molecular interactions between various factors involved in this process (Figure 17).

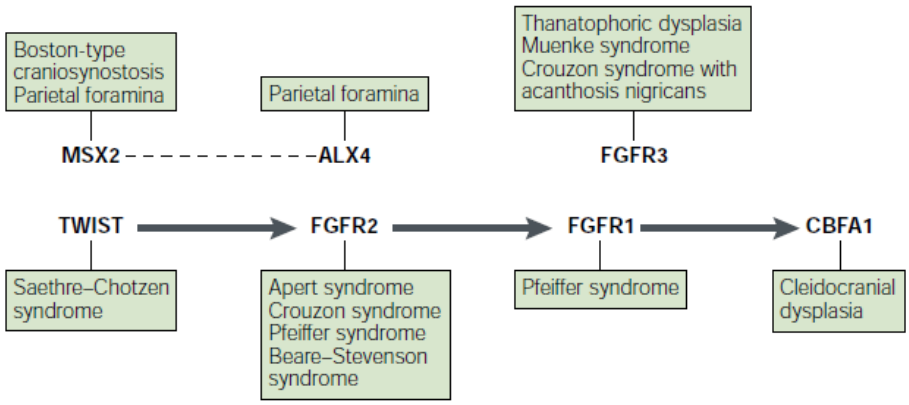


Figure 17. Molecular interactions between several known factors involved in cranial sutures development and malformations, which might develop as a result of various mutations in corresponding genes. Sourced from Wilkie et. al. [79].

Table 1 briefly summarises information on human syndromes, with their respective craniofacial malformations and genes involved. The genes identified in these syndromes were used to search for candidate SNPs in this study, based on the criteria detailed in Section 1.7.

Table 1. Genetic syndromes with various craniofacial abnormalities.

Syndrome	Symptoms	Prevalence	Genetic origin
Aarskog syndrome (OMIM:100050)	Distinct facial features, such as: rounded face, underdeveloped mid-portion of the face (maxilla), small nose with nostrils tipped forward (anteverted), wide-set eyes, crease below the lower lip (hypertelorism)	Rare	Mutations in FGDY1 gene on X chromosome [114]
Alagille syndrome (OMIM: 118450)	Distinct facial features, such as broad forehead, pointed mandible and bulbous tip of the nose and in the fingers	1 in 70,000	Mutations in JAG1 gene [115]
Alfi's Syndrome (OMIM: 158170)	Mental retardation, trigonocephaly, mongoloid eyes, wide flat nasal bridge, anteverted nostrils, long upper lip, cleft lip/palate, short neck, long digits mostly secondary to long middle phalanges	1 in 5 million	Monosomy 9p or 9p22.2-3 deletion [116]
Apert Syndrome (OMIM: 101200)	Various manifestations of craniosynostosis with cleft lip/palate.	Between 1 in 65,000 to 200,000	Mutations in FGFR2 gene [117]
Beckwith-Wiedemann Syndrome (OMIM: 130650)	Characteristic facial appearance and indentations of the ears, a large tongue which may cause breathing, feeding or speech difficulties, one side of the body grows more than the other	Rare	Mutation or deletion of genes H19, KCNQ1OT1 or CDKN1C in 11p15.5 chromosomal region [118-120]
Cohen Syndrome (OMIM: 216550)	Abnormalities of the head, characteristic facial features including high-arched or wave-shaped eyelids, a short philtrum, thick hair, and low hairline	Rare	Mutations in COH1 gene [121]
Cri-du-chat Syndrome (OMIM: 123450). Other name: 5p deletion syndrome	Abnormal larynx and epiglottis which causes a distinct sounding cry. The name literally means "cry of the cat." Other symptoms include mental retardation, small head (microcephaly). Characteristic facial features at birth include a large nasal bridge, round face, wide-spaced eyes, low-set ears, and a down-turned mouth. As the child gets older the facial features change and a long, narrow face is more commonly observed	1 in 50,000 live births	Mutations in two candidate genes: Semaphorine F (SEMA5A) and delta catenin (CTNND2), potentially involved in cerebral development [122]
Crouzon Syndrome (OMIM: 123500)	Craniosynostosis disorder causing secondary alterations of the facial bones and facial structure. Common features include hypertelorism, parrot-	1 in 60,000	Mutations in FGFR2 gene [96, 123]

	beaked nose, short upper lip, hypoplastic maxilla, and a relative mandibular prognathism		
Down Syndrome (OMIM: 190685). Other name: Trisomy 21	People with Down Syndrome have similar facial features including a flattened facial profile, upward slanting eyes, small over-folded ears, flat nose and small mouth with a protruding tongue. They can also have low muscle tone, a shorter than typical neck, a single crease across the palm of the hand, heart defects, and varying levels of intellectual disability	1 in 600-1000 live births. Trisomy 21 is the most common trisomy seen in live born individuals	Extra copy of chromosome 21 in each cell. Each person with Down syndrome may have slightly different symptoms due to variations in chromosomal abnormalities (e.g. Partial or full copy of chromosome 21). Several candidate genes have been identified in Down syndrome critical region, such as DSCR1, DSCR2, DSCR3 and DSCR4 [124] and SHH [125]
Edward Syndrome. Other name: Trisomy 18	Small head (microcephaly), small jaw/mouth (micrognathia), low-set malformed ears, cleft lip/cleft palate, upturned nose, narrow eyelid folds, widely spaced eyes, clenched fists with overlapping fingers, mental retardation, growth deficiency and other skeletal and organ anomalies	1 in 3000-8000 live births. 80% of people with this condition are female	Extra chromosome 18 in each cell. Trisomy 18 is the second most common trisomy seen in live born individuals
Floating-Harbor Syndrome (OMIM: 136140)	Short stature, a triangular shaped face with broad bulbous nose, long eyelashes, deep-set eyes and a wide mouth with thin lips	Rare	Mutations in SRCAP located in 16p11.2 chromosomal region [126]. Rubinstein-Taybi syndrome (OMIM: 180849) shows phenotypic overlap with Floating-Harbor syndrome and is caused by mutation in the CREBBP gene, for which SRCAP is a

			coactivator
Fragile X Syndrome (OMIM:300624)	Range of learning disorders, distinctive facial appearance with large ears and a long face, prominent jaws, speech and language problems		A mutation in the FMR1 gene located on the X chromosome [127, 128]. Within this gene, there is a region containing the sequence "CGG", which is repeated multiple times. Normally the sequence is repeated no more than 55 times in the gene. However, Fragile X Syndrome occurs when a person has more than 200 "CGG" repeats in the FMR1 gene. A person who has more than 55 repeats, but less than 200, is considered a "pre-mutation carrier." These individuals do not have Fragile X Syndrome themselves but are at risk of having children affected with the disorder since the number of repeats could expand in the next generation
Langer-Giedion Syndrome (OMIM:190350(	Short stature, small head, distinctive facial features including deep-set eyes, a bulbous nose, long narrow upper lip and missing teeth	Rare	Deletion of 8q23.2 to q24.1 chromosomal region. Candidate gene in this region: EXT1[129]
Noonan Syndrome (OMIM:163950)	Variable phenotype, which may change with age, many characteristics of which overlap those of the Turner syndrome. Short stature and mild mental	1 in 1,000 to 2,500 live births	Mutation in the PTPN11 gene on chromosome 12q24.1[130, 131]

	retardation are the main features of this syndrome. Characteristic facial features including short webbed neck and low-set posteriorly rotated ears		
Pallister Killian Syndrome (OMIM: 601803)	Coarse face with a high forehead, sparse hair on the scalp, an abnormally wide space between the eyes, a fold of the skin over the inner corner of the eyes and a flat nasal bridge with a highly arched palate	Rare	Mosaicism for tetrasomy of chromosome 12p [132]
Patau Syndrome Other name: Trisomy 13	Common features include: heart defects, small heads (microcephaly), cleft lip and/or palate, small eyes that are close together, extra fingers (polydactyly) and various skeletal abnormalities	1 in 10,000	Trisomy of chromosome 13 [133]
Pfeiffer Syndrome (OMIM: 101600). Other name: Craniofacial-Skeletal-Dermatologic Dysplasia type 1, 2 and 3.	Craniosynostosis, midface deficiency, cloverleaf skull, broad thumbs, broad great toes	1 in 100,000	Mutations in FGFR1, FGFR2 and FGFR3 [95, 134]
Saethre-Chotzen Syndrome (OMIM:101400). Other name: Acrocephalosyndactyly type III	Acrocephaly, asymmetry of the skull, low set hairline, wide and tall forehead, thin, long pointed nose, small low-set ears, cleft palate	1 in 25,000 to 50,000	Mutations in FGFR2 and TWIST1 [90-93]
Smith-Magenis Syndrome (OMIM: 182290)	Abnormalities of the craniofacial area such as brachycephaly, midface hypoplasia, small ears, broad nose and cleft palate. Overlapping features with Potocki-Lupski syndrome	Rare	Mutations in RAI1 gene [135, 136]
Treacher Collins Syndrome (OMIM:154500)	Various craniofacial abnormalities such as antimongoloid slant of the eyes, coloboma of the lid, micrognathia, microtia and other deformity of the ears, hypoplastic zygomatic arches and macrostomia	1 in 25,000 to 50,000	Mutations in TCOF1 gene [137-140]
Turner Syndrome (Monosomy X)	People with Turner Syndrome are females and typically have short stature, a webbed neck, heart defects, swelling of the hands and feet, and	1 out of 2,500 girls	Females with only one X chromosome [108]. Potential involvement of

	characteristic facial features		SHOX gene [141]
Velo-Cardio-Facial Syndrome (OMIM: 192430)	Highly variable phenotype with cleft palate, heart abnormalities, typical faces and over 180 other clinical findings	1 out of 4000 live births	Point mutations in TBX1 [87, 89, 142]
Waardenburg Syndrome (OMIM: 193500)	Characterized by pigmentary abnormalities of the hair, skin, eyes and facial structures, including broad nasal bridge	Rare	Mutations in PAX3 gene [143].

1.5.1.1. Orofacial clefting as a link to the normal facial variation

Next to craniosynostosis, orofacial clefting (OFC) is one of the most common facial birth defects seen in approximately 1 in 600 births. OFC covers a wide range of facial abnormalities, which can be divided into 3 main groups:

- Cleft lip only (CL)
- Cleft palate only (CP)
- Cleft lip with palate (CL/P)

The main mechanism for the OFC is believed to be a failure of palatal fusion during embryonic development of the frontonasal prominence [144-146], as shown in Figure 18.

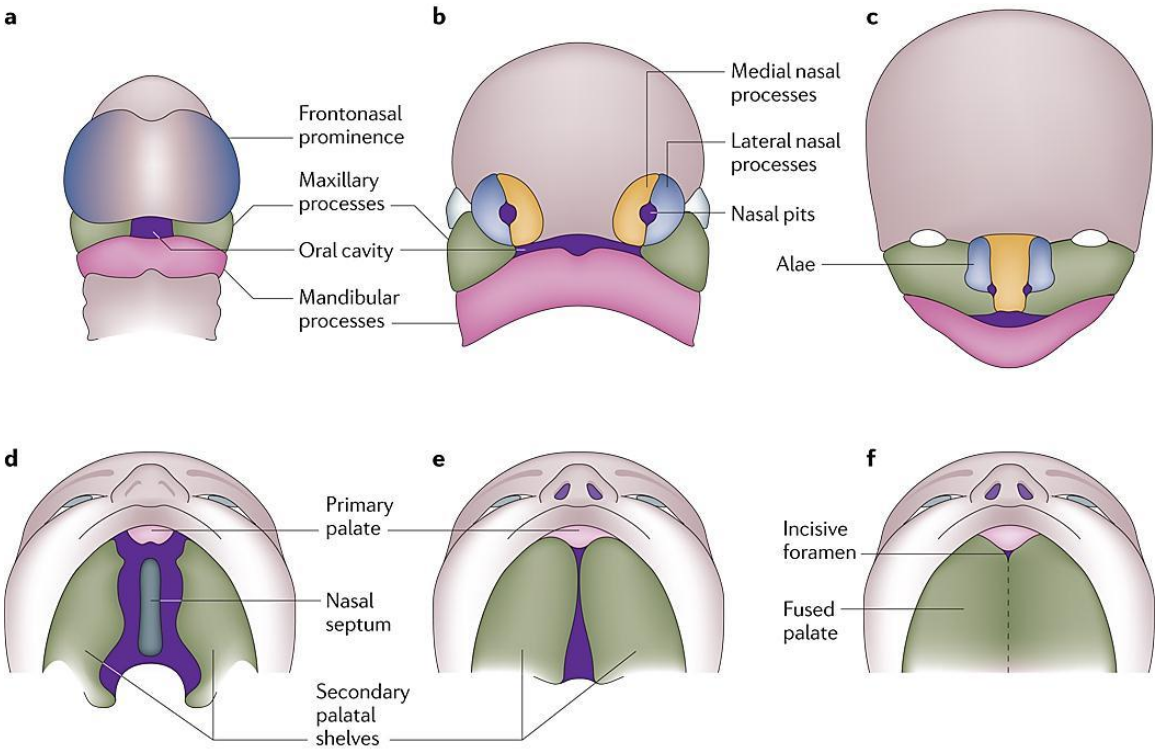


Figure 18. Development of the lip and palate. A) The developing frontonasal prominence, paired maxillary processes and paired mandibular processes surround the primitive oral cavity by the fourth week of embryonic development. B) By the fifth week, the nasal pits have formed, which leads to the formation of the paired medial and lateral nasal processes. C) The medial nasal processes have merged with the maxillary processes to form the upper lip and primary palate by the end of the sixth week. The lateral nasal processes form the nasal alae. Similarly, the mandibular processes fuse to form the lower jaw. D) During the sixth week of embryogenesis, the secondary palate develops as bilateral outgrowths from the maxillary processes, which grow vertically down the side of the tongue. E) Subsequently, the palatal shelves elevate to a horizontal position above the tongue, contact one another and commence fusion. F) Fusion of the palatal shelves ultimately divides the oronasal space into separate oral and nasal cavities. **Sourced from M. Dixon et. al. [145].**

The clefts can be unilateral and bilateral, as shown on Figure 19. OFC can occur as part of a complex craniofacial syndrome or as a non-syndromic facial defect (NSCL/P), with the latter being more common.

While most research has focused on the syndromic form of OFC, a few recent studies have explored NSCL/P in various populations and identified several mutations in candidate genes that may cause this disorder [145, 147-154]. Several studies have shown between 40% - 60% concordance for NSCL/P in monozygotic (MZ) twins and between 5% - 10% in dizygotic twins (DZ), suggesting environmental as well as genetic influences [155, 156]. Interestingly, the highest prevalence of NSCL/P was observed in the Asian population, followed by Caucasians and Africans [155]. Based on the

anthropometric measurements, Asian faces are known to be the widest among the three populations. This fact may reflect the higher NSCL/P prevalence in this population as a matter of failure of correct timing of palatal fusion. A meta-analysis performed on several studies, using cephalometric measurements of unaffected parents of NSCL/P showed a significant increase in several values, including wider interorbital, nasal cavity and upper facial dimensions; narrower cranial vaults; longer cranial bases; longer and more protrusive mandibles; shorter upper faces and longer lower faces and mostly significantly – an increase in facial and nasal cavity width, compared to the control group [157].

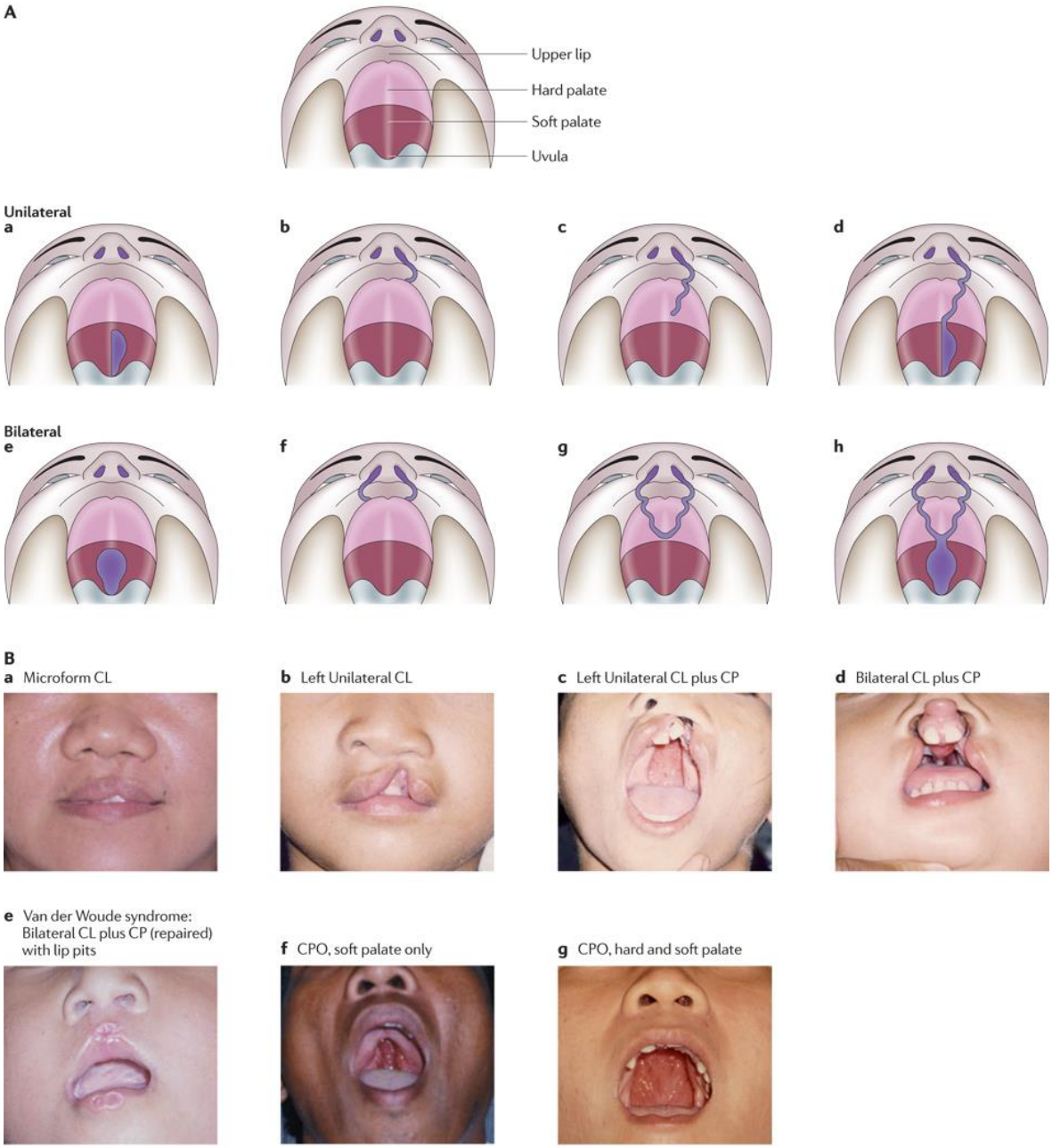


Figure 19. Types of clefts. A) Illustrative drawings of types of cleft lip and/or palate (CLP)114. a) and e) show unilateraland bilateral clefts of the soft palate; b), c) and d) show degrees of unilateral cleft lip and palate; f), g) and h) show degrees of bilateral cleft lip and palate. Clefts are indicated in purple. B) A collection of images of different types of clefts, some with associated anomalies such as lip pits. Descriptions are given above the images. CL, cleft lip; CP, cleft palate; CPO, cleft palate only. **Sourced from M. Dixon et. al. and M. Muenke [145, 158].**

It can therefore be hypothesised that family-based studies of NSCL/P may help to reveal genes that influence normal facial width variation [43, 157, 159]. A recent study indeed found an association between nose width and SNP rs1258763 near the GREM1 gene (cysteine knot superfamily 1) and between bizygomatic distance and SNP rs987525 near the CCDC26 gene (retinoic acid modulator) [160].

Table 2 summarises the candidate genes and respective mutations found to be involved in NSCL/P and such genes were used as candidates in this research project.

Table 2. Summary of genes and known mutations with a role in NSCL/P, which were used as candidate genes for this project.

Gene ID	Top SNP (if known)	Function (if known)
ABCA4	rs560426	Abundant in the retina; expressed at lower levels in the brain. Mutations implicated in disorders of the retinal system.
ARHGAP11A	rs1258763	-
ARHGAP29 (PARG1)		Level of expression might be affected by folic acid.
BMP4		Induces cartilage and bone formation. Also act in mesoderm induction, tooth development and limb formation.
C17orf67		-
CCDC26	rs987525	-
CRISPLD2		-
DGKE		Highly selective for arachidonate-containing species of diacylglycerol (DAG). May terminate signals transmitted through arachidonoyl-DAG or may contribute to the synthesis of phospholipids with defined fatty acid composition.
FGF8		Plays an important role in the regulation of embryonic development, cell proliferation, cell differentiation and cell migration. Required for normal brain, eye, ear and limb development during embryogenesis.
FGFR2		Promotes cell proliferation in keratinocytes and immature osteoblasts, but promotes apoptosis in differentiated osteoblasts.
FMN1		Controls the rate of actin nucleation, thus influencing cell mortality. Fmn1-knockout mice exhibit oligodactylism and show reduced activity of the BMP4 signaling pathway.
FOXE1		This gene functions as a thyroid transcription factor which likely plays a crucial role in thyroid morphogenesis.
GREM1		Member of the BMP antagonist family. Involved in fibroblast growth factor (FGF) signaling of limb bud outgrowth. Acts in cooperation with nog in axial skeleton development.
GSTT1		Glutathione S-transferase (GST) theta 1 is a member of a superfamily of proteins that catalyze the conjugation of reduced glutathione to a variety of electrophilic and hydrophobic compounds.
IRF6	rs642961	Probable DNA-binding transcriptional activator. Key determinant of the keratinocyte proliferation-differentiation switch, involved in appropriate epidermal development. Causative for Van der Woude syndrome.
KCNK18		Outward rectifying potassium channel. Produces rapidly activating outward rectifier K(+) currents. May function as background potassium channel that sets the resting membrane potential.
KIAA1598	rs7078160	Involved in generation of the asymmetric signals required for neuronal polarization and axon outgrowth.
MAFB	rs13041247	Highly expressed in the epithelia of palatal shelves and in the medial edge during the process of palatal fusion in rodents.

MSX1		Act as transcriptional repressor during embryogenesis through interactions with components of the core transcription complex and other homeoproteins.
MTHFR		Catalyzes the conversion of 5,10-methylenetetrahydrofolate to 5-methyltetrahydrofolate, a co-substrate for homocysteine remethylation to methionine.
MYH9		Function as cellular myosin that appears to play a role in cytokinesis, cell shape, and specialized functions such as secretion and capping.
NOG	rs227731	Antagonist of members of the TGF-β superfamily (e.g. BMP4), which are involved in mammalian palatogenesis and missense/nonsense variants are associated with cleft microforms in humans.
None	rs9574565	Located in the intergenic region 13q31.
NTN1		Putative function in axon guidance and cell migration during development.
PAX7	rs742071	Involved in early specification of neural crest stem cells.
PDGFC		Platelet-derived growth factor that plays an essential role in the regulation of embryonic development, cell proliferation, cell migration, survival and chemotaxis.
PIK3R5		Phosphatidylinositol 3-kinases (PI3Ks) phosphorylate the inositol ring of phosphatidylinositol at the 3-prime position and play important roles in cell growth, proliferation, differentiation, motility, survival and intracellular trafficking.
PIK3R6	rs9788972	Phosphoinositide 3-kinase gamma is a lipid kinase that produces the lipid second messenger phosphatidylinositol 3,4,5-trisphosphate. Seems to be involved in angiogenesis.
PVRL1		Encodes a calcium(2+)-independent cell-cell adhesion protein that plays a role in the organization of adherens junctions and tight junctions in epithelial and endothelial cells.
SCG5		Acts as a molecular chaperone for PCSK2/PC2, preventing its premature activation in the regulated secretory pathway. Plays a role in regulating pituitary hormone secretion.
SUMO1		Ubiquitin-like protein that can be covalently attached to proteins as a monomer or a lysine-linked polymer.
TGFA		Encodes a growth factor that is a ligand for the epidermal growth factor receptor, which activates a signaling pathway for cell proliferation, differentiation and development.
TGFB3		Encodes a member of the TGF-beta family of proteins, which is involved in embryogenesis and cell differentiation.
THADA		-
VAX1		Transcription factor that may function in dorsoventral specification of the forebrain. Required for axon guidance and major tract formation in the developing forebrain.
ZFP36L2	rs7590268	Probable regulatory protein involved in regulating the response to growth factors.

1.5.2. Animal models in craniofacial disorders research

Animal models have been extensively used to generate craniofacial syndrome models. This has led to the identification of genetic factors that cause these conditions in humans and allowed a better understanding of the embryonic craniofacial development. These animal models include mouse, bird and fish species, dog breeds and primates [161].

1.5.3. Mice and fish models

Targeting of specific genes in mice has generated more than 90 loss-of-function mutants that have various craniofacial malformations [62, 162-164]. Interestingly, some of these mutants demonstrate variation in facial morphology that can be considered normal, rather than abnormal, for example short or long nose, prominent or depressed face and different shapes of ears (illustrated in Figure 20). These and other mouse models are available from Jackson mouse laboratory [164].

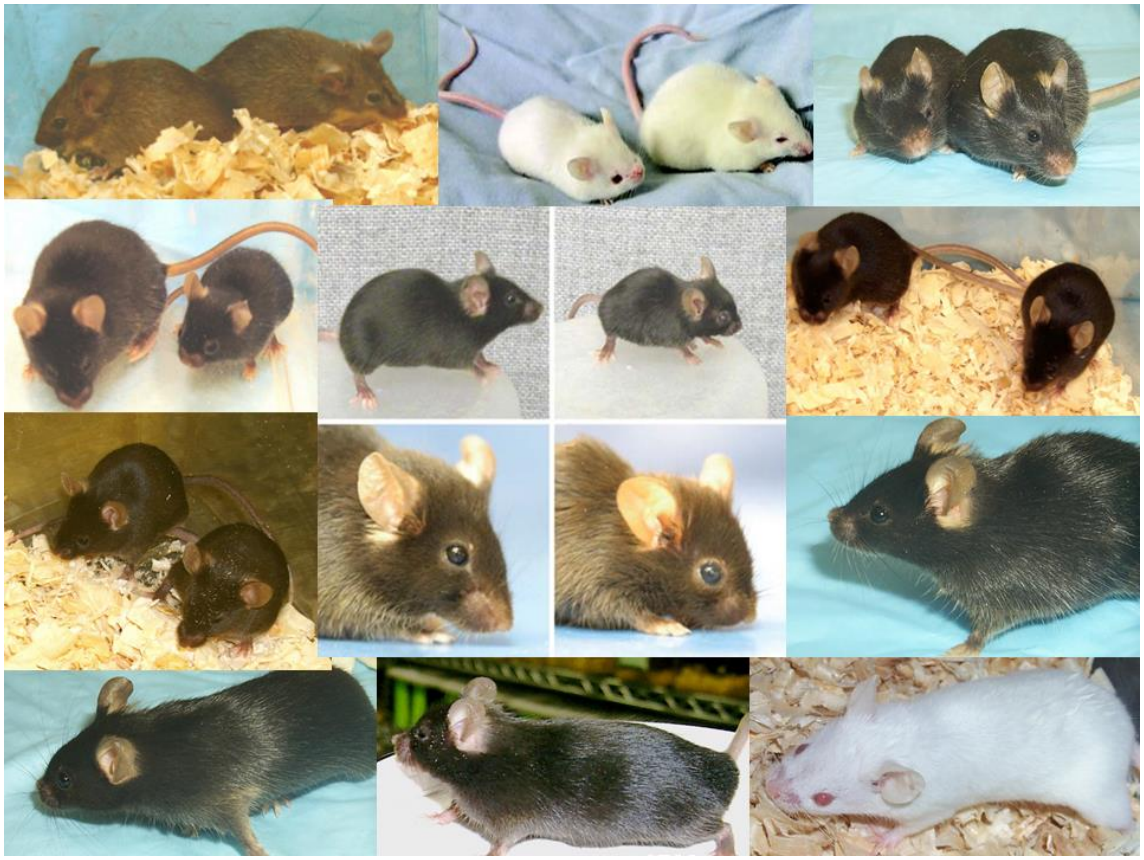


Figure 20. Examples of mouse mutants showing various craniofacial defects. Reproduced from Jackson Laboratory web site (http://www.informatics.jax.org/).

Despite the successful identification of craniofacial genes in animal models, the generation of knockout alleles does not always provide a clear picture of the function of a particular gene in craniofacial development. There are two main reasons for this. Firstly, the disruption of a natural function might be lethal if a gene plays a critical role in earlier developmental stages. Genes included in this category encode various transcription factors, mostly belonging to homeobox-containing genes and represent a major portion of genes believed to be involved in craniofacial embryogenesis. Secondly, the expression pattern of the targeted knockout gene may not be fully visible due to genetic redundancy. These synergistic interactions might be detected by producing double-homozygous mutants [165]. Secondly, the phenotypic effects on a system of interest might be masked by a major effect on another system that developed earlier. For example, a role for a particular gene involved in skull vault formation will not be detected if its loss also affects neural-tube formation.

Research on craniofacial mutants in fish and particularly in zebra fish (*Danio rerio*) [166-171] and cichlid species [172, 173] has also revealed a number of genes involved in various stages of craniofacial embryogenesis. Relatively easy mutagenesis and fast generation turnover in fish provided numerous craniofacial skeleton mutants with genes orthologous to other vertebrates, including humans. Many of the affected genes identified in fish replicated those identified in mouse, while some were novel candidates. The genes, identified in both animal models (listed according to the associated craniofacial malformations) include:

- Neural tube defects, involving Pax3, Twist, Gli3, Dlx5, Tcfap2a and Cart1 genes [90-93, 167, 168, 174-179].
- Neural crest defects, involving Dlx, Msx, Pax, Prrx gene families, as well as Gsc and Hoxa2 genes [165, 180-191].
- Cranial truncations, involving Shh, Pcsk6, Sil, Lhx1, Hesx1, Otx2 genes [149, 168, 171, 192-201].
- Defects of sensory organs, involving Pax2, Pax6, Chx10, Chrd, Rax and Bmp7 genes [168, 188, 189, 202-204].
- Abnormalities of skeletal differentiation, involving Runx2, Col2a1, Col11a1, Crt11 and Hspg2 genes [98, 100, 167, 205-208].
- Clefts of the secondary palate, involving Jag2, Tgfb3 and Lhx8 genes [209-212].

A recent study discovered that craniofacial morphology in mice is regulated by distant – acting transcriptional enhancers [213]. Hundreds of such enhancer sequences were

identified, while targeted knockout of three of them resulted in slight alteration of the cranial shape in mice. The authors hypothesise that sequence and/or copy number variation in these regions may contribute to the variance in craniofacial appearance in human populations.

In addition to the mouse and zebra fish, other animal models have been studied to explore additional factors involved in craniofacial development in these organisms and potentially in humans.

1.5.4. Dog breeds

Dogs have been domesticated for more than 30,000 years [214, 215] and maybe as early as about 135,000 years ago [214, 215]. On the other hand, a recent study have found that dogs and wolves diverged only 11,000–16,000 years ago from a common ancestor in a process involving extensive admixture [216]. In any respect, centuries of selective breeding has resulted in approximately 400 different breeds, according to International Kennel Association (Fédération Cynologique Internationale), making dogs probably the most diverse domesticated animal. During this process, dogs' cranial morphologies and their brains, have been significantly affected [217-220]. This makes the dog a useful model organism to explore craniofacial variation and its potential implications in humans. Several studies have explored the difference between various dog breeds, especially between the brachycephalous or "shortened head" breeds, such as bulldog, pug, boxer and the dolichocephalous or "elongated head", such as greyhound, saluki, collie breeds [137, 139, 218, 221-224].

While most research has focused on the genetics of the skull morphology, rather than the soft tissue, a few studies have revealed genes associated with additional phenotypic traits [225, 226]:

- The MITF gene, associated with white spotting in bull terriers and boxers;
- a 133kb duplication of the chromosomal region containing 5 genes: FGF3, FGF4, FGF19, ORAOV1 and CCDN1, associated with the "ridge" trait in ridgeback dogs;
- The HAS2 gene, associated with excessive skin wrinkling in shar-pei dogs;

The genes that have been found to be involved in influencing the craniofacial diversity in dogs and considered candidate genes for this project include: RUNX2, TCOF1,

BMP3, MSX2, THSB2 and SMOC2. A more comprehensive list of candidate craniofacial genes summarized from available studies in different animal models is detailed in Supplemental Table S1.

1.5.5. Avian species

Additional valuable information for determining genes involved in craniofacial development was contributed by Brugmann et. al. [209], who studied transcription factor (TF) gene expression patterns in cranial neural crest cells in the developing beaks of ducks, quails and chicken. This work demonstrated a species-specific transcription factor expression profile of 232 genes in neural crest cells that precedes morphological differences between the species. The most dramatic changes between species were found in the Wnt signalling pathway, with a 20-fold up-regulation of Dkk2, Fzd1 and Wnt1 in the duck compared with the other two species. Twenty-two genes of the differentially expressed genes, including Fgfr2, Jagged2, Msx2, Satb2 and Tgfb3, have previously been shown to be involved in a variety of mammalian craniofacial defects. Seventy-two of the differentially expressed genes represented new loci, potentially involved in human craniofacial disorders, making them candidate genes affecting normal craniofacial appearance (Table 3).

Table 3. Example of 72 genes differentially expressed between developing beak of the chicken, quail and duck. Based Brigmann et.al., (2010).

Syndrome name	Differentially expressed genes in region
Van Der Woude syndrome	GJB5, MYCBP, PTCH2
Rosselli-Gulienetti syndrome	OCT11
Smith-Magenis syndrome, Potocki-Lupski syndrome, Van Der Woude syndrome, Cleft palate	ALDH3A2, SREBF1
Hypercalciuria	ASCL1, MORF4
Coloboma	BC052625, CDK5, PAXIP1L
Chromosome 2p16.1-P15 Deletion syndrome	BTF3L2
Seckel syndrome	CDKN3, TRIM9
Hemifacial Microsomia	CRIP1, JAG2, TRIP11
Orofacial Cleft	ARC, DKK2, FGF2, FLJ12517, LOC152485, MADH1, MGC15631, NKX6A, NR3C2, RAI15, RELB, PITX2, PRDM5, RREB1, TGFB2
Craniofacioskeletal syndrome	FGF13
Chromosome 1q43-Q44 deletion syndrome	FLJ12517
Microphthalmia, Armfield X-Linked mental retardation syndrome	FMR2
Craniometaphyseal Dysplasia	FOXO3A, SCML4
Ectrodactyly, Ectodermal Dysplasia, Cleft Lip/Palate syndrome 1	FZD1
Chromosome 2q32-Q33 Deletion syndrome	FZD5, FZD7, NAB1, SATB2
Seckel Syndrome 2	GATA6
Pierre Robin syndrome	GPRC5C
Chromosome 3q29 Microdeletion syndrome	HES1
Orofacial Cleft, Adrenoleukodystrophy, Zellweger syndrome	HES5
Miller-Dieker Lissencephaly syndrome	HIC1
Moebius syndrome	HMG1
Cdags syndrome	HRIHFB2122, LOC90322, MFNG, PPARA, SOX10, TCF20
Split-Hand/Foot malformation	HSPC063, TLX1, NFE2L2
Chromosome 1p36 Deletion syndrome	ID3
Hypothalamic Hamartoma	IGFBP3
Prader-Willi syndrome	KLF13
Cat Eye Syndrome, Digeorge syndrome, Opitz Gibb syndrome,	LZTR1, PCQAP
Tarp syndrome	MLLT7
Craniosynostosis, Adelaide Type, Wolf-Hirschhorn syndrome	NKX1-1, WHSC1
Holoprosencephaly	PAX9, PFKL, PTTG1IP, TNRC15
Axenfeld-Rieger syndrome type 2	RGC32
Digeorge Syndrome/Velocardiofacial syndrome	TAF3
Zellweger Syndrome	THBS3
Syndactyly Type I	WNT6

1.5.6. Primates

Primates are the closest human relatives, although information about the genetics of their craniofacial development is limited. A study on the craniofacial structure of baboons examined 43 anthropometric measurements on each of the 830 skulls and revealed 14 significant quantitative trait loci (QTLs) for 12 craniofacial traits [227]. A subsequent study on 981 human cephalograms, identified additional QTLs for 10 cranial measurements [228]. The candidate genes in QTLs included BMP6, WNT1, WNT10B, WNT5A, which have previously been shown to be involved in craniofacial embryogenesis in other model organisms.

Supplemental Tables S1 and S2 provide a comprehensive list of candidate genes, which were used in this project.

1.6. Genomic factors that may influence craniofacial phenotype variation

In addition to SNPs, other sequence variations and epigenetic factors may potentially affect facial appearance. Epigenetic differences can regulate transcription and subsequent gene expression by DNA methylation, histone modifications and chromatin packaging, without affecting the DNA sequence directly. On the other hand, SNPs in regulatory elements (such as enhancers) may also trans-regulate various factors, which affect gene expression [213]. This type of regulation may be responsible for phenotypic differences between monozygotic twins, which become more visible with age [38, 57, 58, 60, 156, 229]. While these epigenetic modifications may also play a role in determining craniofacial morphology, this research project focuses on single nucleotide polymorphisms.

1.6.1. DNA methylation pattern

DNA methylation is known to down regulate or silence genes. In vertebrates it is catalysed by the enzyme DNA methyltransferase, resulting in conversion of cytosine to 5-methylcytosine and typically occurs at CpG sites. The human DNA has about 80-90% of CpG sites methylated, except for CpG islands, which are regions of DNA comprising approximately 65% unmethylated CG residues. CpG islands are found in the promoters of approximately 56% of mammalian genes, including all ubiquitously expressed genes. Thus the methylation pattern has an important role in regulation of the transcription activity and gene expression. DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, suppression of repetitive elements and carcinogenesis. It has been hypothesized that epigenetic differences of this kind may affect the difference in phenotypic appearance of monozygotic twins [58]. This type of genetic polymorphisms may also play a role in affecting craniofacial morphology.

1.6.2. Histones modifications

Histone acetylation

There are several types of histone modifications, but the most extensively studied is acetylation. Histone acetylation is associated with chromatin's "open state" and resulting transcription activation [230]. Acetylation, performed by a number of histone acetyltransferases, neutralizes a positive charge on histones via the negative charge of the amine group on the histone tail, thus decreasing the affinity between the DNA and specific histone or the whole nucleosome (formed by eight histone sub-units and 147 bases of DNA). This event makes DNA more accessible to various transcription factors, regulating gene expression. This type of nucleosome remodelling may have a more global "house-keeping" effect at the cellular level and is not always inherited [231].

Histone methylation

This type of histone modification occurs by addition of methyl groups to arginine and lysine amino acids in histone proteins. Histone methylation can alter transcription by both repression and activation. In addition, methylation of a specific lysine residue on the histone tail can make it a target for continuous acetylation and deacetylation [232].

A recent study suggests that heterochromatin silencing by histone methylation may involve RNA interference and may be heritable [233]. Another recent study proposes, that a particular type of histone demethylase regulates zebrafish brain and craniofacial development [169]. These findings support a model in which PHF8 demethylase regulates zebrafish neuronal cell survival and jaw development in part by directly regulating the expression of the homeodomain transcription factor MSX1/MSXB [180]. Based on these and other sources, PHF8 and MSX1 were chosen as putative candidate genes for this research project.

A recent histone-code hypothesis speculates that modified histones play a more active role, recruiting other proteins via specific recognition of protein domains, rather than just altering the interaction with a DNA molecule [234, 235]. The histone code believed to be extremely complex, as each of the four histone pairs, forming a nucleosome can be modified at multiple sites through multiples types of modifications, providing various

levels of transcriptional regulation. For example, given that histone H3 which contains 19 lysines, known to be methylated by up to four methyl groups, can potentially form up to 4^{19} different lysine patterns. This number should be multiplied by additional possible modifications in three arginine methylation sites, nine acetylation sites and at least eight phosphorylation sites in each of the 44 million H3 histones in the human genome [236]. This additional level of regulation may potentially be involved in many cellular processes, including "fine tuning" of the craniofacial phenotype.

1.6.3. Copy number variations (CNV)

CNVs are alterations of genomic DNA that correspond to relatively large regions of the genome that have been deleted or amplified in the genome. CNVs can be caused by genomic rearrangements such as deletions, duplications, inversions, and translocations. This variation accounts for approximately 12% of human genomic DNA and may range from a few kilobases to several megabases in size. It is estimated that approximately 0.4% of the genomes of unrelated people typically differ with respect to copy number [237, 238]. De novo CNVs have also been observed between identical twins, who were believed until recently to have identical genomes [57, 229]. CNVs are less likely to be involved in determining a normal phenotypes, as such large aberrations in the genetic code can significantly alter various protein functions and as a result lead to disease.

1.6.4. Non-coding RNA interference (RNAi)

RNA interference is the process by which RNA molecule interferes with the accumulation of homologous transcripts (mRNA) from genes of the same origin [239]. This sub-class of non-coding RNA (usually 20-25 bases in length) together with other micro-RNAs, may represent a still-hidden layer of signals that regulate various levels of gene expression in physiology and development, transcriptional activity, epigenetic memory through chromatin remodelling, RNA splicing, editing and turnover [240-243]. Each of these networks, separately or together, may play a significant role in normal and abnormal genetic variation of human complex characteristics.

1.6.5. Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms represent one of the most common type of variation in the human genome. SNP markers are typically bi-allelic base substitutions, insertions or deletions that occur approximately every 1 in 100-300 bases along the 3 billion bases of human genome, according to the latest statistics from the NCBI dbSNP database [244]. The HapMap project, one of the largest efforts to identify human variation, has already genotyped approximately 3.2 million tag SNPs, representing 25-35% of common SNPs, in several populations [245]. Tag SNPs are markers in high linkage disequilibrium, which represent a large number of SNPs (haplogroup) that tend to be inherited together. Genotyping of a relatively small number of tag SNPs is more cost-effective and may provide information on other markers, located within the same haploblocks.

Another project, called "1000 genomes", has to date genotyped 15 million single nucleotide polymorphisms, 1 million short insertions and deletions, and 20,000 structural variants in 2,500 samples from 27 populations [246].

In spite of the relatively limited forensic use of SNPs at present, these biomarkers have shown potential to be particularly useful for typing limited and highly degraded forensic samples [3, 247] and predicting pigmentation traits and ancestry [9, 13, 14, 248-252].

1.7. Types of craniofacial candidate SNPs and their selecting criteria

In general, SNPs may be located in coding and/or non-coding regions of genes or in the intergenic regions. Thus, SNPs may be divided into several categories, according to their location and potential function (Figure 21):

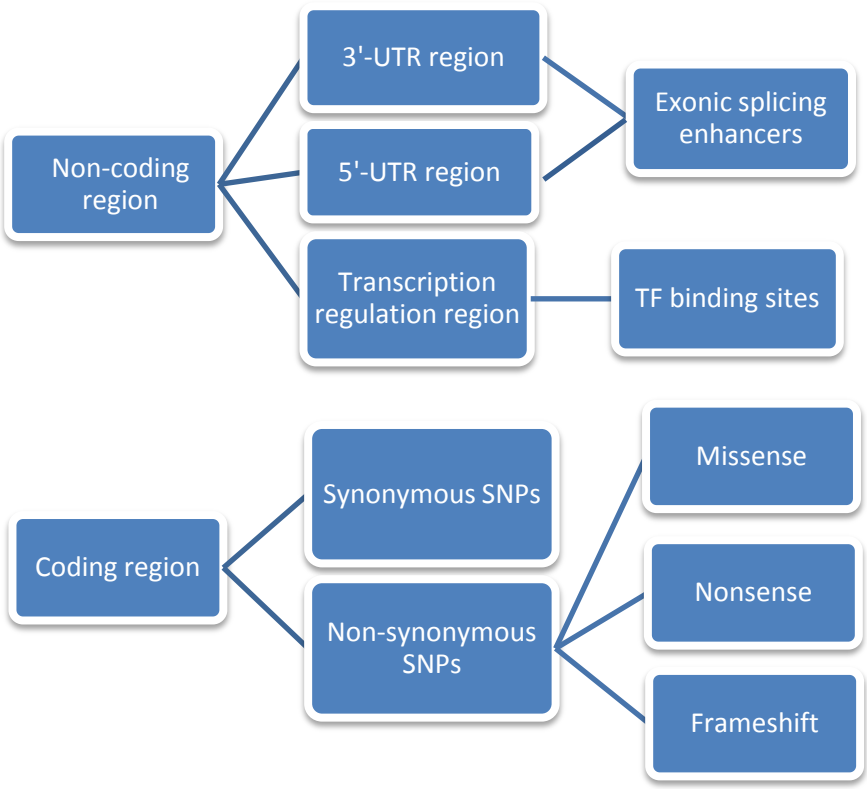


Figure 21. Schematic representation of SNPs classification.

Based on the categorisation shown in Figure 21 and given that particular facial features are more prevalent in specific populations, it can be hypothesised that markers belonging to the following categories could have an impact on the proposed craniofacial candidate genes function:

- ✓ Ancestry informative markers (AIMs) - high population differentiation markers with Fst >0.5

  Genetic differentiation between two populations occurs when they diverge and become significantly isolated to limit gene flow between them. Population differentiation is sensitive to a variety of evolutionary events and can be measured by both Fst and LD-based tests, such as LRH, iHS and XP-EHH [253-256]. However, ancient selective events can mainly be detected with Fst-based tests [257]. In humans, the starting point of population differentiation is believed to coincide with the exodus from Africa, some 50,000 to 75,000 years ago [255, 258, 259]. Thus the Fst value can reveal the effects of natural selection for the last 75,000 years. Another advantage of using the Fst value to detect population diversity is that, unlike other tests, Fst is SNP specific, allowing targeting of specific genetic variants under selection [260]. A recent study has shown that a positive as well as negative selection has guaranteed the regional adaptation of human populations by increasing population differentiation primarily at nonsynonymous and 5'-UTR variants [257]. Based on the HapMap Phase II data, Barreiro et.al. identified a list of 582 genes showing signs of positive selection (Fst>0.65) and containing at least one polymorphism. Another study of 26,530 SNPs identified 174 candidate genes with SNPs that were a target of positive selection and as a result show high Fst values [261]. Several complementary studies have been published recently, providing a valuable resource for identification of AIMs in candidate genes for craniofacial morphology [256, 258, 262-268]. It was hypothesized that high Fst markers should provide a primary targeted SNPs subclass for the current project.

  An example of such two SNPs, showing different distribution in the world –wide population is illustrated in Figure 22. Hypothetically, a 'T' allele in SNP rs12595883 could be associated with a long face or more prominent nose (more common in the European population), while an ancestral 'C' allele is associated with a short face or depressed nose. Similarly, an ancestral allele 'A' in SNP rs12598094 could be associated with the decrease in facial width (eu-eu distance), while the 'C' allele is associated with the increase in this distance (more common in the Asian population, as shown in this figure). This is just a simplified illustration, as the variation in the same facial trait in different populations could be affected by different SNPs.

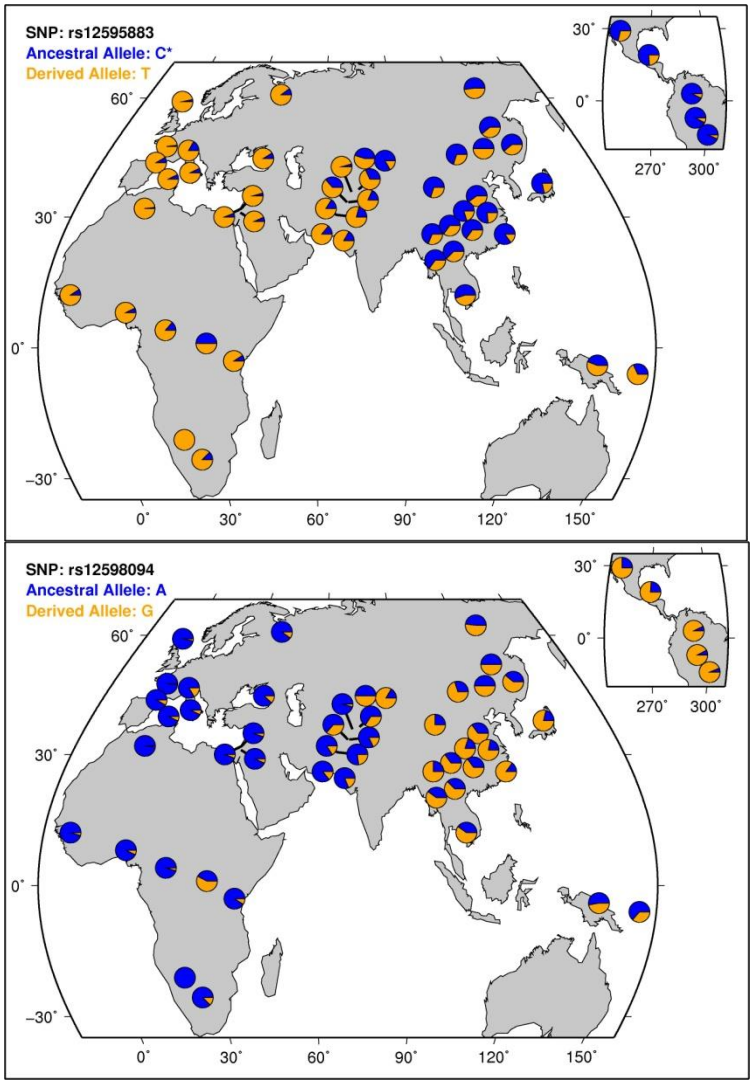


Figure 22. Example of two SNPs from dbSNP database with various population distribution. Sourced from NCBI website (http://www.ncbi.nlm.nih.gov/projects/SNP/).

✓ Non-synonymous SNPs (nsSNPs) in coding regions

The polymorphisms located in the coding regions can be of two types: synonymous and non-synonymous. The latter can impact the translation process by either missense, nonsense or frameshift mutations. According to a recent study on the basis of phase II HapMap data, there are approximately 15,259 nsSNPs out of 2,841,354 SNPs tested in the human genome [257]. Any of these polymorphisms may result in a change to the polypeptide sequence, thus possibly influencing the protein folding and function and affecting the phenotypic expression of the gene. Such a significant change may have a subsequent dramatic effect on the gene expression, leading to various disease

conditions (supplemental Tables 1S and 2S) or simply normal variation in craniofacial phenotype.

Due to the redundancy of the genetic code, synonymous SNPs typically do not alter protein structure and function. However, recent studies show that these mutations, previously believed to be "silent", may affect mRNA stability and alter protein function and as a result lead to disease [269-273]. Nevertheless, this research has focused mainly on nsSNPs, as having the major impact on potential phenotype.

- ✓ SNPs in splice sites

Recent studies show that up to 94% of human genes are regulated by alternative splicing, providing a major mechanism for transcript diversity [274]. Alternative splicing is a complex process, regulated by a variety of cis and trans acting factors, many of which recognise donor and acceptor splicing sites [275]. Mutations in splice sites can directly affect exon configuration by disrupting the splicing process. Therefore, SNPs located in splice sites can have a more pronounced effect on protein structure than non-synonymous SNPs.

- ✓ SNPs in regulatory regions

DNA transcription represents one of the key steps in transforming the genetic code into phenotype and is regulated through various sequence elements, such a promoters, enhancers and silencers [276].

Although our understanding of human transcriptional regulatory sequences is still limited, it is obvious that polymorphic variation in these regions may alter the level of protein expression and, as a result, play a significant role in phenotype diversity in general and craniofacial appearance in particular. This hypothesis has been successfully supported by a recent study in mice [213].

- ✓ Tag SNPs

There are at least 12 million SNPs in the human genome [245, 246, 277]. Genotyping all these markers would be time consuming and costly. It is known that SNPs which are located in close proximity, tend to be inherited together.

Thus, genotyping such a set of strongly associated SNPs (a haplotype) can provide cost-effective information about genetic variation in a particular region of interest.

The tag SNPs are chosen based on the haplotype frequencies in different populations. The HapMap project has identified about 3.2 million tag SNPs, which should provide almost as much information as the 10 million SNPs in the human genome. The use of tag SNPs should significantly reduce the number of SNPs needed for genotyping in this project. Those SNPs will be derived from markers from any of the categories described above.

1.8. Current state of the forensic DNA analysis and potential applications of this project

1.8.1. Forensic application of the short tandem repeats (STRs) and mitochondrial DNA

Most forensic laboratories use autosomal short tandem repeats (STRs) as the primary technology for routine identification purposes. Short tandem repeats represent a sub-class of microsatellites, containing between 2-5 bp repeats. The forensic DNA community uses tetra-nucleotide repeats, as they can be amplified with less artefacts than bi- or tri-nucleotide repeats. PCR amplification of STR markers usually results in amplicons ranging from 100 bp to 500 bp and more recently of shorter amplicons (Mini-STRs), which may vary between 51 bp to 211bp [278, 279]. This technology is fast, robust, sensitive and uses a widely adopted capillary electrophoresis platform as a genotyping method. Most commercial STR kits provide multiplex reactions of at least 10-20 markers, which can amplify as little as 0.3-0.5 ng of DNA.

Y chromosome STR typing and in some instances, mitochondrial sequencing of the HV1 and HV2 regions are also used. The use of these methods include a resolution of female-male mixtures with a very low portion of male fraction (e.g. DNA under fingernails) and historical paternity issues (e.g. Thomas Jefferson slave child) by Y-STR typing [280], as well as successfully dealing with very degraded DNA evidence (e.g. mass disasters) [281] or maternal ancestry cases (e.g. Romanov family remains identification) by mitochondrial sequencing [282].

1.8.2. Forensically relevant SNP classes

There are two main routes for forensic use of SNPs, namely identification (including lineage-informative) and intelligence (ancestry and phenotype informative). The following section outlines briefly each SNP category.

Identity testing SNPs

Markers in this category are chosen according to the relatively high average heterozygosities of >0.4 and low Fst values of <0.06. The combination of approximately 50 of such SNPs is highly discriminatory and the probability of any two unrelated individuals having identical DNA profile is extremely low [283-285].

SNP-based identification assays have been developed for forensic purposes on numerous platforms [284, 286-289]. Efficient multiplex reactions for the amplification and typing portions of an assay can be optimized for the small sized amplicons, allowing more genetic data to be gleaned from an already limited sample. The selection of the typing platform is typically driven by the number of samples to be typed, as well as the number of SNPs to be interrogated. Although these assays are quite powerful in individualizing an evidentiary sample, they do not yield any information regarding phenotype appearance (such as pigmentation or facial features information) to narrow the search for suspects [284, 289].

Lineage informative SNPs

Lineage informative SNPs are sets of tightly linked SNPs that function as haplotype markers to identify missing persons or mass disaster victims through kinship analyses. These markers have a low mutation rate, are inherited as a haploblock and are therefore better than STRs for analysis of kinship cases, especially if the evidence and reference samples are separated by several generations, as in the Romanov Royal family identification case and president Thomas Jefferson's slave child [280, 282].

Additionally, autosomal SNP haploblocks, which are inherited together and provide a higher discrimination power than individual SNPs inside this block, can also be used [3, 290].

Ancestry Informative SNPs (AIMs)

AIMs are used in order to establish a reliable probability of an individual's biogeographical ancestry. SNPs in this category are distributed with different frequencies in the world populations and therefore have low heterozygosity and high Fst

values [249, 291], similar to phenotype informative markers discussed below. Genotyping these markers may reveal ancestry information of the person and indirectly infer some phenotypic characteristics of investigative value [267]. An additional application of AIM markers might be detection of risk factors for various diseases and developing efficient treatment in the field of personalized medicine [265, 268]. These markers are discussed in more detail in Chapter 4.

Phenotype informative SNPs

This category of forensic DNA markers enables prediction of the physical appearance of an individual and in some aspects is overlapping with AIMs. These markers aim to detect particular phenotypic characteristics such as skin, hair or eye colour from a DNA sample for investigative purposes. Some of these traits can be also predicted indirectly knowing the bio-geographic ancestry of a person (although not with the same accuracy).

In the past few years, many studies focused on the detection of polymorphisms in pigmentation genes, have been published [5, 248, 292, 293] and a few forensic assays have been developed, for detecting hair [5, 294, 295] and iris colour [292, 293] or both together with ancestry [12, 292, 296, 297].

The genetic basis of additional phenotypic characteristics, such as height, weight and age is still poorly understood. However, the number of articles on the genetics of these complex traits is growing, providing hope for developing forensic assays for predicting these features in the not too distant future [298-300].

1.8.3. Advantages and limitations of identity-informative SNPs over STRs in forensic DNA analysis

SNPs have several advantages over STRs:

- SNPs are the most common form of variation at the DNA level, representing approximately 85% of all DNA variation [2, 3, 301, 302].
- SNPs are more stable over time. Their mutation rates in general are much lower (down to 10^{-8}) than those of STRs (most frequently 10^{-3}), making some of these markers appropriate for kinship or biogeographical analysis [278, 303].
- Compared to the current semi-automation with STRs, genotyping technologies for SNPs are amenable for full automation with efficient high throughput and cost-effectiveness [285, 289, 304, 305].
- Due to much smaller amplicons, genotyping of severely degraded DNA would have greater success with SNPs as compared to STRs [283, 306, 307].
- SNPs may provide information on externally visible characteristics of the source of the DNA sample, not available through STR markers.

Possible limitations of identity SNPs in forensic DNA analysis include:

- Due to bi-allelic nature of most SNPs, their statistical power of discrimination is lower than of STR commercial kits. As a result, more markers are needed for a high discriminating power.
- The analysis of mixed DNA profiles and mixture interpretation is more complicated with SNPs, due to difficulty to resolve different contributors alleles.
- All current forensic DNA databases are based on STR data and include millions of profiles, making it problematic to replace them with SNPs.

SNPs are unlikely to replace STRs as a primary method of forensic DNA identification [278, 303]. However, they may serve as a complementary tool in attempts to identify highly degraded human remains and a valuable tool in prediction of phenotypic and ethnic characteristics for investigative leads. The ultimate goal is to integrate the various categories of SNPs in one forensic assay, which would eventually provide information on identity, ancestry, pigmentation, appearance and other externally visible traits of the person, who is the source of the DNA sample.

1.9. Bioinformatical web-based resources for SNP search

In the last decade substantial progress in genomic studies, especially in Genome Wide Association Studies (GWAS), has resulted in the availability of a growing number of bioinformatics resources for candidate marker selection and evaluation. These resources provide mostly free access to a range of data on gene functions and interactions, SNP locations, population data and statistical information. Given the size of the human genome, the number of polymorphisms and possible genomic and protein interactions, the amount of information offered by these resources is overwhelming. However, many of these tools are not systematized and not updated on an ongoing basis, which creates difficulties in candidate genes and SNPs selection process.

One of the first successful attempts to systematize currently available resources for candidate markers selection was the "GeneEpi toolbox" [308]. This resource offers various workflows and resources for SNPs selection and subsequent evaluation (Figure 23).

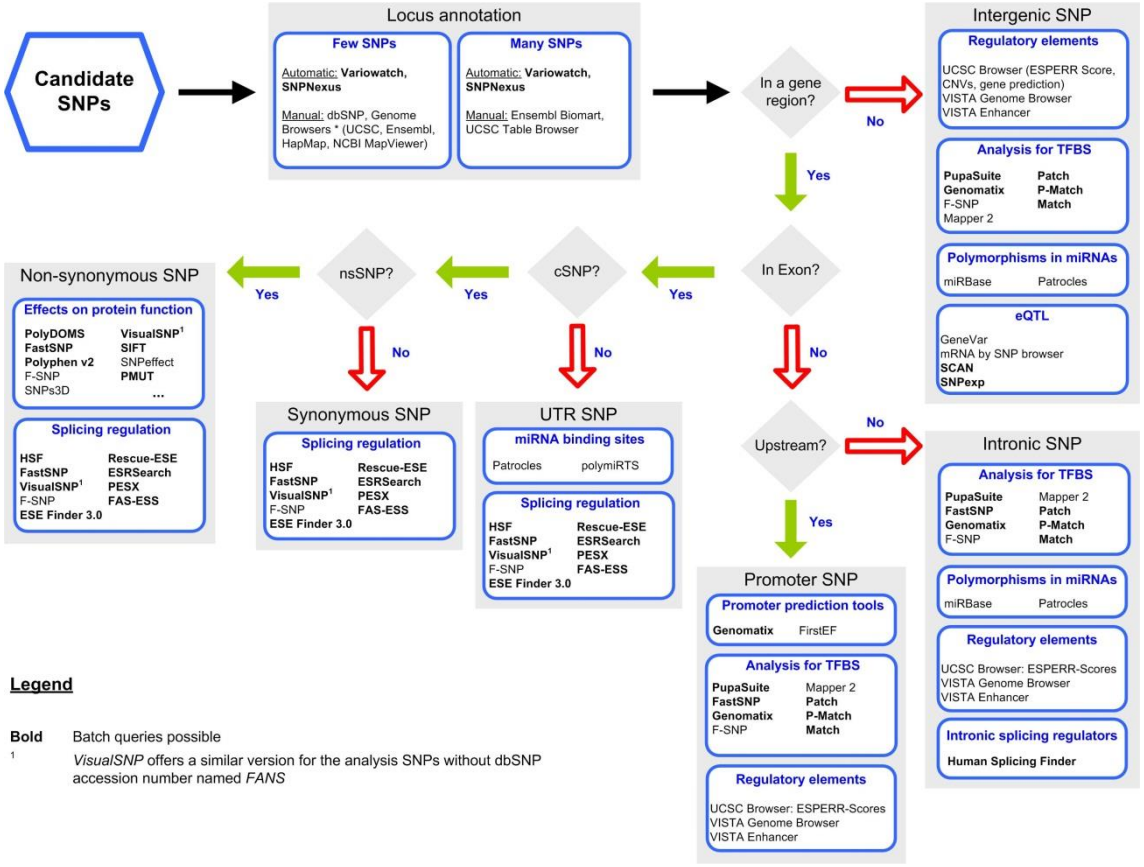


Figure 23. GeneEpi toolbox schematic representation.

The major resource for searching SNP-related information is the dbSNP database. This database represents an international central repository for single base nucleotide substitutions, short deletions and insertion polymorphisms [244]. It represents a part of a greater database, offering a large amount of information on gene interactions, protein functions, taxonomy and disease-related genetic information (Figure 24). Additional web resources that were used in this project are detailed in the Chapter 2.

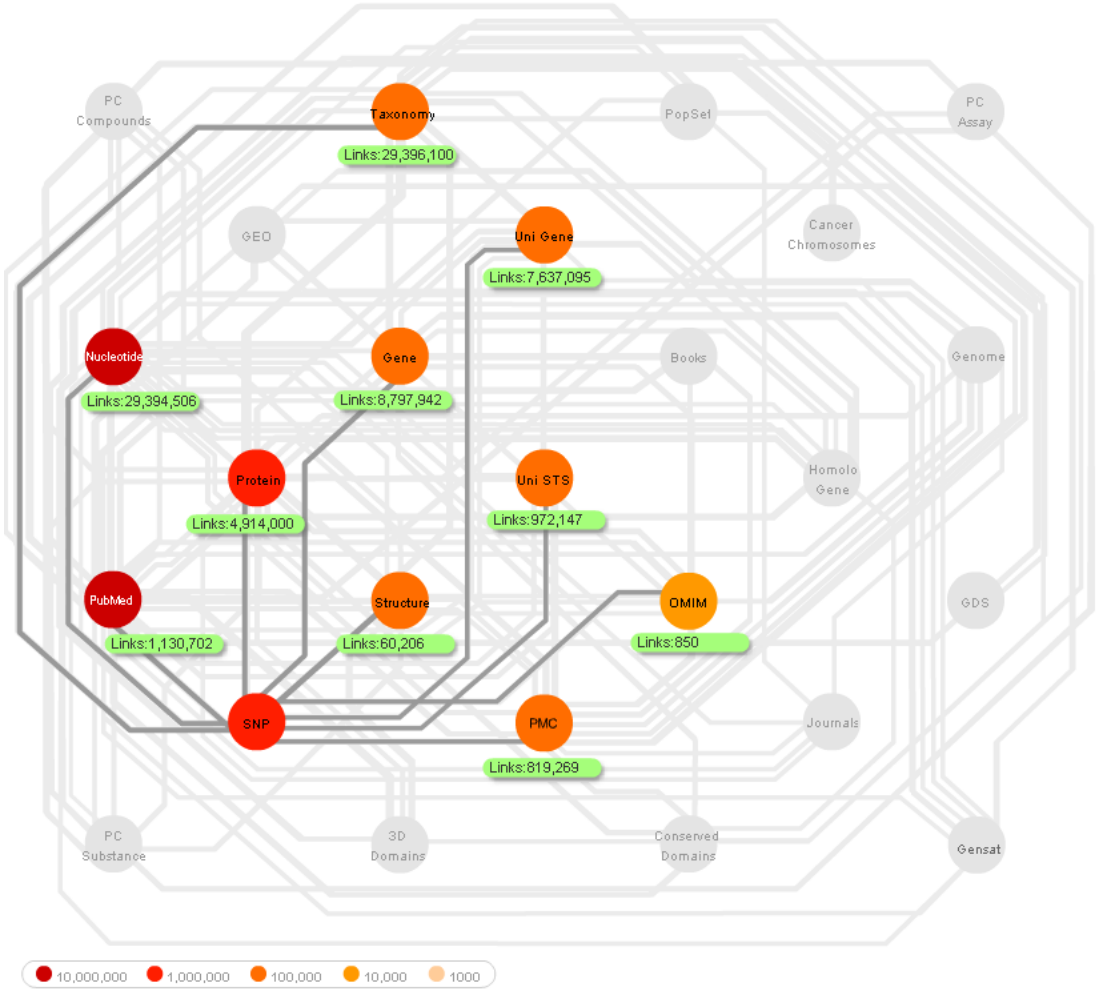


Figure 24. Model of Entrez databases, showing interactions among them. The dbSNP database consists of over 12 million SNPs (different species) up to January 2011.

1.10. Targeted massively parallel sequencing as a novel forensic genotyping platform

There are many different SNP genotyping technologies which can be classified by the type of molecular mechanism employed: primer extension, allele specific hybridization, sequencing, oligonucleotide ligation and invasive cleavage, reviewed by Sobrino et. al. [304]. The major SNP detection methods include fluorescence, luminescence, mass spectrometry and more recently pH change detection [309-311].

While the majority of forensic DNA laboratories use PCR, followed by capillary electrophoresis (CE) for the separation of STRs fragments [2, 3], alternative DNA typing methods utilize Massively Parallel Sequencing (MPS), also known as Next Generation Sequencing (MPS). MPS is a relatively new method, which uses various platforms and chemistries to perform DNA sequencing in an unprecedentedly fast manner [312]. Current forensic DNA analysis targets only tens of STR markers, while MPS methods can simultaneously type thousands of markers from the same sample, including STRs, SNPs, INDELs and mitochondrial DNA polymorphisms in a single multiplex reaction. As MPS becomes more established in basic research, it is also being investigated as an alternative method to capillary electrophoresis for forensic DNA analysis [313-315]. This approach would allow recovery of maximum useful genomic information from a limited DNA sample, without the need for additional amplification reactions. Most MPS methods enable processing of multiple samples through barcoding, providing increased output compared to CE methods. MPS platforms allow simultaneous analysis of millions of individual DNA reads, which would allow better resolution of complex mixtures [316]. An additional advantage of sequencing is the ability to detect intra-STR polymorphisms, which cannot be detected by CE, as it is a length-based detection method. This can maximize the amount of information revealed from a DNA sample, providing better resolution of genomic data [317]. It is possible that in the next few years, commercial forensic DNA assays will offer a "one for all assay", able to generate not only the STR and SNP-based identity-informative data, but also type ancestry–informative, lineage–informative and phenotype – informative markers. The current major disadvantages of the MPS technology are the high cost and relatively complicated and time-consuming data analysis. These pitfalls however, are likely to be solved with the continuous progress in this technology.

One of the relatively recent MPS technologies is the Ion Torrent^{TM} platform, manufactured by Life Technologies$^{TM.}$. This methodology is based on the detection of

hydrogen ions, which are released during DNA synthesis by incorporation of complementary dNTP molecules (A, G, T or C). In this type of sequencing, DNA polymerization occurs directly on an electronic chip, which incorporates more than six million micro wells embodying a template and a sensor. During the sequencing, each ion that is released in the solution, causes a change in the electric signal (a change in pH), which is detected by a semiconductor sensor (a highly sensitive pH meter) and transferred to the computer for real-time data analysis. The main advantage of this technology is that signal detection does not involve labelled nucleotides, optics or any conversion process, but occurs directly through electronics. The Ion Torrent platform provides a relatively rapid sequencing time and low cost, compared to alternative MPS platforms. These advantages played a major role in choosing this platform for this research project. The main limitations of this system include the relatively short reads (up to 400 bp) and high error rate in sequencing of homopolymer DNA repeats [318, 319].

One of the examples of an Ion Torrent-based targeted MPS assay, which has the potential to be used in forensic DNA typing is the Ampliseq™ assay, which provides several SNP panels for targeted DNA sequencing and detection of specific mutations (e.g. Cancer panel). This approach offers remarkable multiplexing capabilities, simultaneously amplifying thousands of custom markers from a relatively low amount of DNA template (10 ng versus approximately 250 ng needed for other platforms) in relatively short period of time.

There are several consecutive steps in the custom Ampliseq protocol (illustrated in Figure 25). These steps include:

- **Primer design.**

  Designing and building a custom panel of targeted markers, which is submitted to Life Technologies™ for primer design. The current primer pipeline success rate is subject to multiple problems with primer design in specific genomic locations (e.g. homopolymer repeats). The panel may include thousands of markers, amplified in one or two pools.

- **Library construction.**

  This step consists of amplification, primer digestion, barcode ligation and purification steps. It is performed using 10ng of DNA template with the possibility of processing multiples samples, using up to 96 specific barcodes. The final

sequencing library produced by generating DNA fragments flanked by the Ion Torrent sequencing adapters. This step is time and labour consuming, if performed on multiple samples.

- **Template preparation.**

  This step involves clonal amplification of the library fragments on Ion Sphere particles by emulsion PCR and is performed using an automated instrument – One Touch™. Subsequently, the Ion Sphere particles coated with template are deposited on a chip and sequenced using the Personal Genome Machine (PGM).

- **Data analysis.**

  Data generated through the sequencing run are simultaneously transferred and analysed on the Ion Torrent server. The initial analysis includes signal processing and base calling, producing DNA sequences associated with individual reads. Following alignment of the sequenced data against reference sequence, the sequence can be subsequently transferred to third-party software (e.g. Ion Reporter™) for annotation of specific variants.

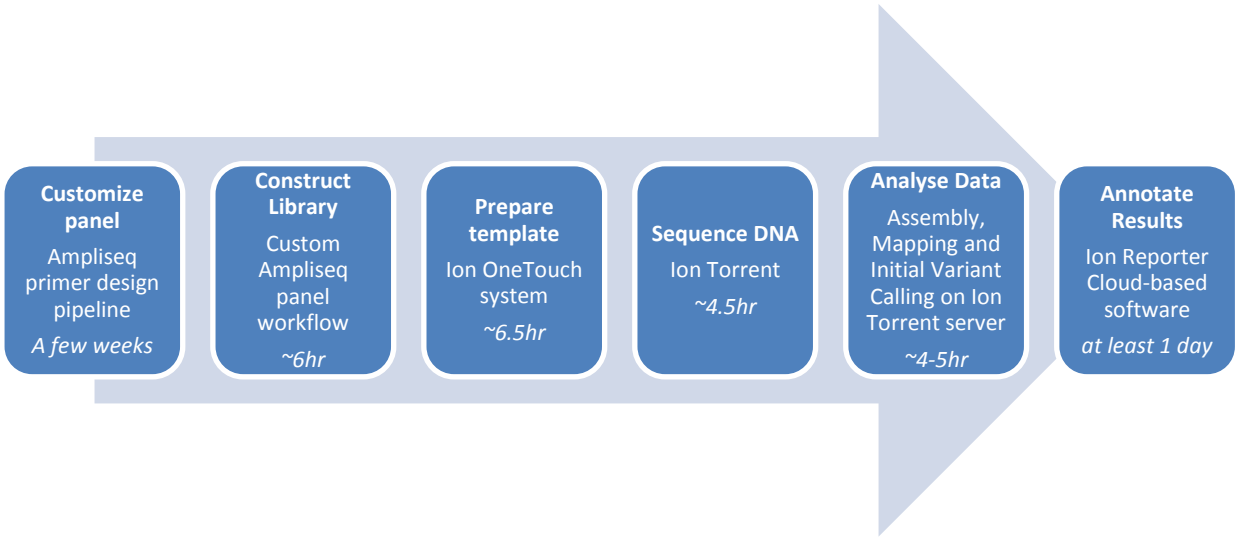


Figure 25. A general illustration of the custom Ampliseq protocol.

Although the actual sequencing step takes only a few hours to perform (on the IonTorrent platform), the library preparation and the data analysis steps are currently the major limiting factors of this process and may take days to weeks. Sequencing of hundreds and possibly thousands of markers from numerous samples would generate a large amount of data, which needs to be analysed in a user-friendly and time-efficient way. Additionally, the accuracy of sequencing can be severely affected by a number of factors, such as DNA quantity and quality, efficiency of library and template preparation as well as sequence alignment algorithms. Current bioinformatical platforms provide only a partial solution for these problems, still insufficient for forensic market needs, which requires standardized protocols and very high accuracy of genotyping from a variety of challenging samples. Nevertheless, given that MPS and related bioinformatics is a very dynamic field, it should be expected that solutions for these problems would emerge in the near future.

1.11. Project aims

The main hypothesis of this research project is that SNPs in candidate genes may influence normal craniofacial variation in humans. The major aim of this research is to identify a set of SNPs potentially involved in normal human craniofacial variation, determine which of these markers are associated with craniofacial variation and subsequently test the power of these SNPs to predict visual appearance of a person. A secondary aim is to incorporate additional, previously identified markers, such as identity, lineage, ancestry and pigmentation – informative SNPs and test their prediction power on blind samples. This comprehensive assay would provide maximum information on the source of the DNA sample from a crime scene, helping with facial reconstruction in missing person and mass disaster cases as well as enriching our knowledge of the craniofacial embryogenetics.

The specific aims of this project are summarized below:

AIM 1: To review the literature and relevant web resources in order to create a list of candidate genes and SNPs, potentially involved in normal craniofacial appearance. To generate a shortlist of SNPs, targeted for subsequent genotyping by sequencing.

AIM 2: To collect a database of 300-500 DNA samples, along with 3D facial images and relevant phenotypic information. To analyse the 3D images for a set of craniofacial landmarks and various facial measurements.

AIM 3: To validate and optimize the various methods used in this project, such as DNA extraction, genotyping and 3D image analysis.

AIM 4: To genotype the DNA samples, using a MPS platform and analyse the data for potential association between the craniofacial measurements and additional phenotypic traits, such as pigmentation and ancestry, using statistical software. Following the association study results, to evaluate the statistically significant markers for their prediction power of specific traits.

Chapter 2

Materials and Methods

2.1. Ethics approval

The ethics approval for this project was granted by Bond University Human Ethics committee (RO 510). Most of the volunteers who participated in this study were Bond University students. All volunteers completed a questionnaire and gave informed consent (supplement documents S3 and S4).

2.2. Samples

In total, 587 DNA samples were used for genotyping purposes. This sample set included 534 volunteers who donated both DNA samples and 3D facial images and 14 DNA samples without 3D images. Each participant donated four (4) buccal swabs. The samples without 3D images were accompanied with pigmentation and ancestry information. The mean age of the volunteers was 27.4 year old. The phenotype-related information in questionaries and all the craniofacial measurements were taken by the author. The ethnicity information of the volunteers' grandparents was recorded. Samples of volunteers who had experienced severe facial injury and/or undergone facial surgery (e.g. nose or chin plastics) were not used in the craniofacial traits association study, although some were used for analysing association of markers with pigmentation traits and ancestry.

An additional 39 DNA samples were obtained from an in-house laboratory database, which has pigmentation and ancestry as well as partial cranial measurements information (mostly eu-eu, g-op and cephalic indexes) recorded over a period of 10 years (2000-2009).

All extracted DNA samples were stored at -80^0 C. The 3D facial image database and recorded personal information were stored in a single user access database on a dedicated computer.

An additional set of samples was used for genotyping as a part of a validation study using the GoldenGate assay on the BeadExpress instrument (Illumina). The samples for this study originated from three sources:

- 552 DNA samples from the in-house (Bond University) database collected over a period of 10 years (2000-2009), with pigmentation (eye, hair and skin colour) and ancestry information recorded.

- 365 DNA samples from the Health Science Center DNA database (University of North Texas, Fort Worth, USA). The samples were from the three major US population groups: Caucasian, African American and Hispanic. The samples of the African American ancestry were assigned with dark skin and black hair phenotype.
- 35 DNA samples were sourced from sub-optimal buccal swabs and environmentally challenged fieldwork samples.

A more detailed information on these samples is available in Chapters 3.5.1 and 3.5.2

2.3. DNA extraction

DNA from the samples collected for this study was purified from saliva using one of the following extraction methods. The majority of the samples were extracted manually, using the Isohelix DDK isolation kit (Cell Projects, Kent, UK) as per manufacturer recommendations [320]. Alternatively, samples (n=120) were purified on the EZ1 extraction robot using a semi-automated EZ1 Buccal swabs card (Qiagen, Hilden, Germany). Twenty four additional DNA samples were extracted in duplicate using a DNA IQ purification kit (Promega, US) or a Miniprep kit (Qiagen), according to manufacturer recommendations [321, 322]. Each two (2) of four (4) collected buccal swabs were used for a single DNA extraction.

2.4. DNA quantification

DNA samples were quantified using one of two methods. The majority of the samples were quantified using a Real Time quantitative PCR (q-PCR) method. This assay amplified a 63bp region of the OCA locus. The primer sequences were 5'-GCTGCAGGAGTCAGAAGGTT-3' (forward primer) and 5'-CATTTGGCGAGCAGAATCC-3' (reverse primer) at a final concentration of 200mM. The assay was performed on either Rotor-Gene 6000 (Qiagen), Bio-Rad CFX96 (Bio-Rad, Gladesville, Australia) or 7500 Real-Time (Life Technology) thermal cyclers in a 25µL reaction volume using SensiMix HRM Master Mix (Bioline).

The three-step qPCR protocol consisted of an initial 15 minute 95°C Taq DNA polymerase activation step, followed by 40 cycles of 15 seconds of denaturation (95°C), 10 seconds of annealing (60°C) and 10 seconds extension (72°C). High Molecular

Weight (HMW) human genomic male DNA of known concentration (Promega, Madison, WI) was used as a qPCR quantification standard (0.0254 - 25.4ng/µL). Standard curves with good linearity (R^2 values above 0.99) were accepted for analysis. No template controls (NTCs) were included to monitor contamination during qPCR.

All DNA samples were additionally quantified using the Qubit fluorometer (Life Technologies, Mulgrave, VIC, Australia) prior to library construction as per manufacturer recommendations.

2.5. Candidate genes and SNPs search using bioinformatics resources

The following web resources were used for identification of candidate genes, which may play a role in the embryonic craniofacial development in model organisms and be responsible for normal facial variation. Some web resources were used for several aspects of the candidate marker selection process and are repeated in different categories.

- Search for candidate genes in human and in model organisms
  - Mouse Genome Informatics (The Jackson Laboratory): http://www.informatics.jax.org/
  - dbSNP database: http://www.ncbi.nlm.nih.gov/projects/SNP/
  - GeneCards server [323]: www.genecards.org
  - Human Osteogenesis PCR Array: http://www.sabiosciences.com/rt_pcr_product/HTML/PAHS-026A.html
  - Ensemble genome browser: http://www.ensembl.org/Homo_sapiens/Info/Index
  - HGDP selection browser: http://hgdp.uchicago.edu/cgi-bin/gbrowse/HGDP/
  - AmiGo – the gene ontology database [324]: http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

The following list summarises web resources, used for screening of high population differentiation markers in candidate genes:

- High Fst SNPs and AIMs selection
  - ENGINES: http://spsmart.cesga.es/engines.php?dataSet=engines [325]
  - FstSNP-HapMap3 database of 115,213 SNPs based on HapMap data http://FstSNP-hapmap3.googlecode.com/[326]
  - A map of human genome variation from population-scale sequencing [246]
  - A High-Density SNP Map for Signatures of Natural Selection [261]
  - A list of 582 genes containing at least one genic mutation showing signs of positive selection [327]
  - Database of 5,000 AIMs http://www.cs.rpi.edu/~drinep/HGDPAIMS/ [265]:
  - UCSC genome browser: http://genome.ucsc.edu/cgi-bin/hgTracks?org=human
  - Haplotter: http://haplotter.uchicago.edu/
  - SNP evolution genome browser http://124.16.129.22/cgi-bin/gbrowse/evolution_B36/ [328]:
  - SPSMart database [329]: http://spsmart.cesga.es/

The candidate genes and SNPs list was further screened for additional markers according to the following categories:

- nsSNPs in coding regions
  - International HapMap project [277] http://hapmap.ncbi.nlm.nih.gov/index.html.en
  - 1000 genomes project http://browser.1000genomes.org/ [246]
  - PolyPhen server for prediction of functional effect of human nsSNPs www.bork.embl-heidelberg.de/Poly Phen [330]
  - GeneCards server [323] www.genecards.org

- SNPs in transcription regulation sites selection
  - GeneCards server www.genecards.org [323]
  - F-SNP database: http://compbio.cs.queensu.ca/F-SNP/
  - GRAIL software for examination of relationships between genes: http://www.broadinstitute.org/mpg/grail
  - Multi-genome database of positions and patterns of elements of regulation: http://genome.ufl.edu/mapper/run
  - GWAS catalogue http://www.genome.gov/gwastudies/ [71]:
  - is-rSNP software [331]: http://bioinformatics.research.nicta.com.au/software/is-rsnp/
  - SNPNexus http://www.snp-nexus.org/ [332]:
  - Japanese SNP database: http://snp.ims.u-tokyo.ac.jp/
  - TFSearch: http://www.cbrc.jp/research/db/TFSEARCH.html

- Potentially functional SNPs selection
  - SNP function prediction portal: http://snpinfo.niehs.nih.gov/snpinfo/snpfunc.htm
  - Potentially functional SNP search engine [333] http://pfs.nus.edu.sg/%28S%28e5qybiumwhvbajcdnpyup32q%29%29/QueryInterface_V5_2.aspx
  - Genomic Regions Enrichment of Annotations Tool (GREAT) web-based platform (http://bejerano.stanford.edu/great/public/html)

- Tag SNPs selection
  - Haploview software for haplotype analysis http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview [334]:
  - Snagger http://snagger.sourceforge.net/ [334-337]
  - Tagger : http://www.broadinstitute.org/mpg/tagger/
  - SNPPicker http://www.mybiosoftware.com/population-genetics/4430[338]:
  - LD Tag SNP selection server: http://snpinfo.niehs.nih.gov/snpinfo/snptag.htm

2.6. Target selection and primer design process

Approximately 1,200 markers in the craniofacial candidate genes were selected using various web resources were selected as a list of SNPs with their respective rs numbers and chromosomal location. Approximately 700 additional markers, previously shown to be associated with pigmentation traits, ancestry and identity informative markers were selected from the relevant literature and added to the final markers list.

The final candidate markers list was submitted online at: https://www.ampliseq.com/browse.action for the custom Ampliseq primer design pipeline. Based on the design output (failure of some markers), the marker list was resubmitted with alternative markers, showing high linkage with the markers that failed initial primer design. The final custom Ampliseq primer multiplex set was designed as two separate pools of approximately 850 primer pairs each.

2.7. 3D facial scanning procedure

Craniofacial scans were taken with a Konica Minolta Vivid 910 3-D digitiser. A Minolta medium range lens with focal length of 14.5 mm was used for surface registration. This camera emits an eye safe class I laser (FDA) λ=690 nm at 30 mW with an object to scanner distance of 600-2500 mm and scan time of approximately 2.5 seconds in fine mode. Vivid V910 uses a one-half-frame transfer charged couple device (CCD) and can acquire 307,000 data points. The scanner output data are 640 x 480 pixels for 3D and RGB data. The output data were recorded on a laptop computer with Intel Core2 Duo, 3GHz processor, equipped with Polygone® software. Two daylight fluorescent sources (3400K/5400K colour temperature) were mounted in the room. These lights were placed approximately 1.5 meters from the subject's head, with the main halogen light dimmed, (resulting in a more ambient light coverage and a better image quality).

The scanner was mounted at a distance of approximately 1 meter from the volunteer's head. Each volunteer remained in an upright natural position during the scan. Subjects with long hair pulled their hair behind the ears. Glasses and earrings were removed.

Each volunteer was scanned three times from different angles (front and two sides). Three cranial measurements (Eu-Eu, G-Op, V-Gn) were taken, using a digital spreading calliper (Paleo-Tech Concepts, USA).

The scanned images were registered and aligned using overlapping coordinates (Figures 26 and 27). The final merged 3D image was produced by semi-automatically aligning all scans and deleting non-overlapping or unnecessary data (e.g. neck area, hair and jewellery). The complete coordinates of each merged 3D image were saved in a vivid file format (.vvd).

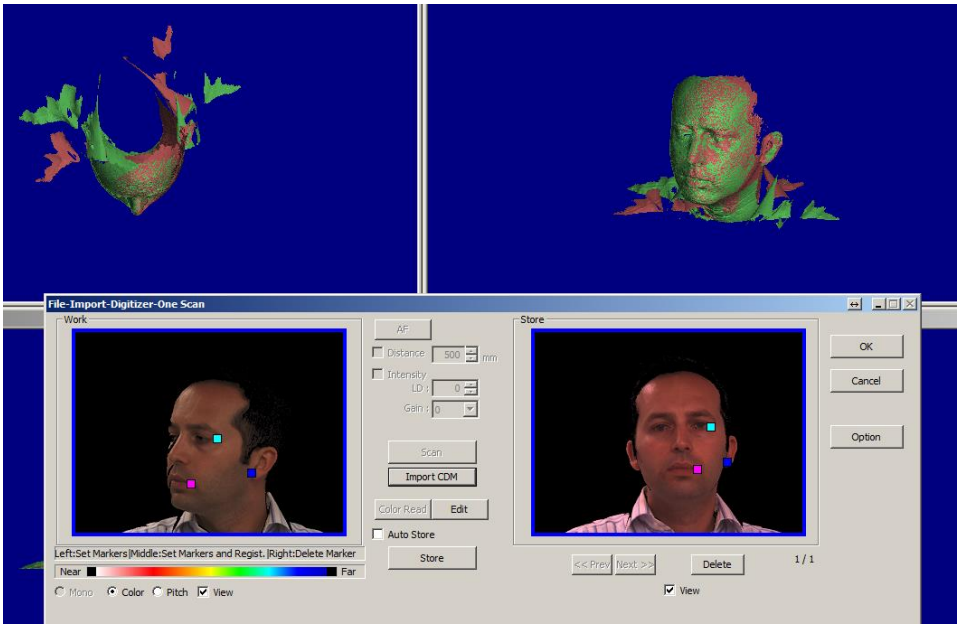


Figure 26. An illustration of alligning of two facial scans with the Polygon software.

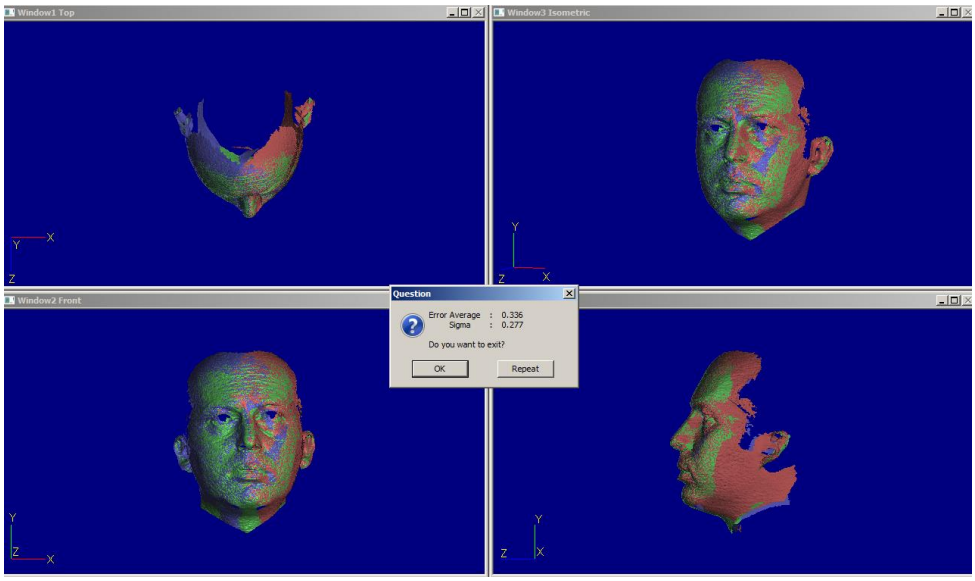


Figure 27. An illustration of the final 3D image, produced by merging three facial scans.

The image file (see an example on Figures 28) was transferred to Geomagic® software for initial image manipulation as below:

- Three .vvd merged objects (representing three facial scans) from Polygone® software were imported and saved in .wrp format as one merged object.
- Compensation for the noise error made by the scanner was performed by automatically moving points to statistically correct locations with fixed deviation limit of 0.03mm.
- A "Mesh doctor" function was applied on low quality images. This procedure automatically repaired imperfections in the polygon mesh of the scanned object.
- Some selected holes, not covered by the scanning process and required for analysis (such as eyes or nostrils area) were filled in the polygon object.
- The resulted 3D merged image was saved for further location of the craniofacial landmarks.

Figure 28a

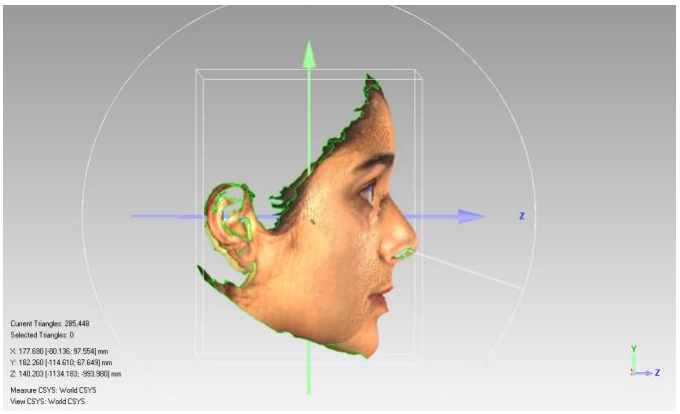


Figure 28b

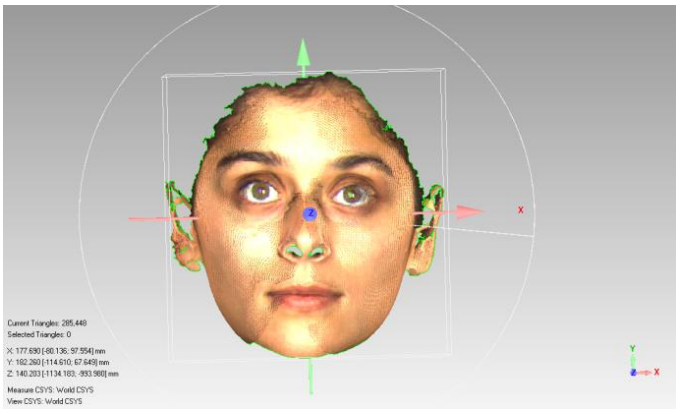


Figure 28c

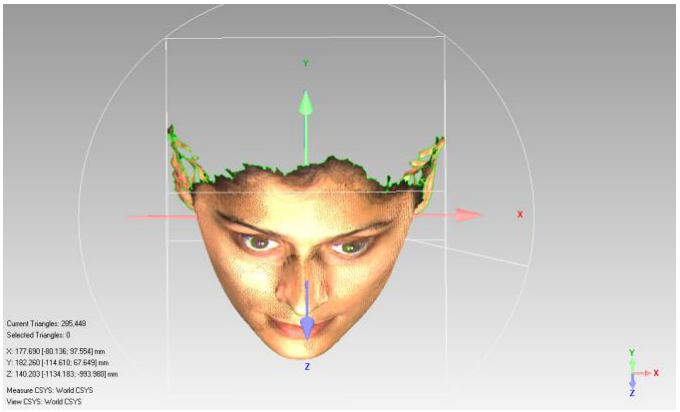


Figure 28. An illustration of a merged image output after initial image processing in the Geomagic software. a) side view, b) front view, c) top view.

2.8. Cranial measurements

Three cranial measurements were performed using a sliding digital calliper (Paleo-Tech Concepts, USA). The measurements were as follows:

- V-Gn (Craniofacial height)
- Eu-Eu (Head Width)
- G-Op (Head Length)

Based on craniofacial and body height measurements, three craniofacial ratios were calculated using a Microsoft Excel software:

- Cephalic index – (eu-eu)*100/(g-op)
- Head width – Craniofacial height index (eu-eu)*100/(v-gn)
- Head – Body height index (v-gn)*100/(body height)

2.9. Facial measurements

Facial measurements were recorded upon location of specific anthropometrical landmarks and included linear and angular distances as well as ratios between these

distances. The following section describes the landmarking protocol used and provides details regarding the craniofacial measurements obtained.

2.9.1. Landmarking protocol

Each 3D image was analysed to allocate 32 facial landmarks using Geomagic software. Each landmark was represented by ‘x’, ‘y’ and ‘z’ coordinates as part of the Cartesian coordinate system (Figure 29). The coordinates were exported to an Excel spread sheet for subsequent analysis.

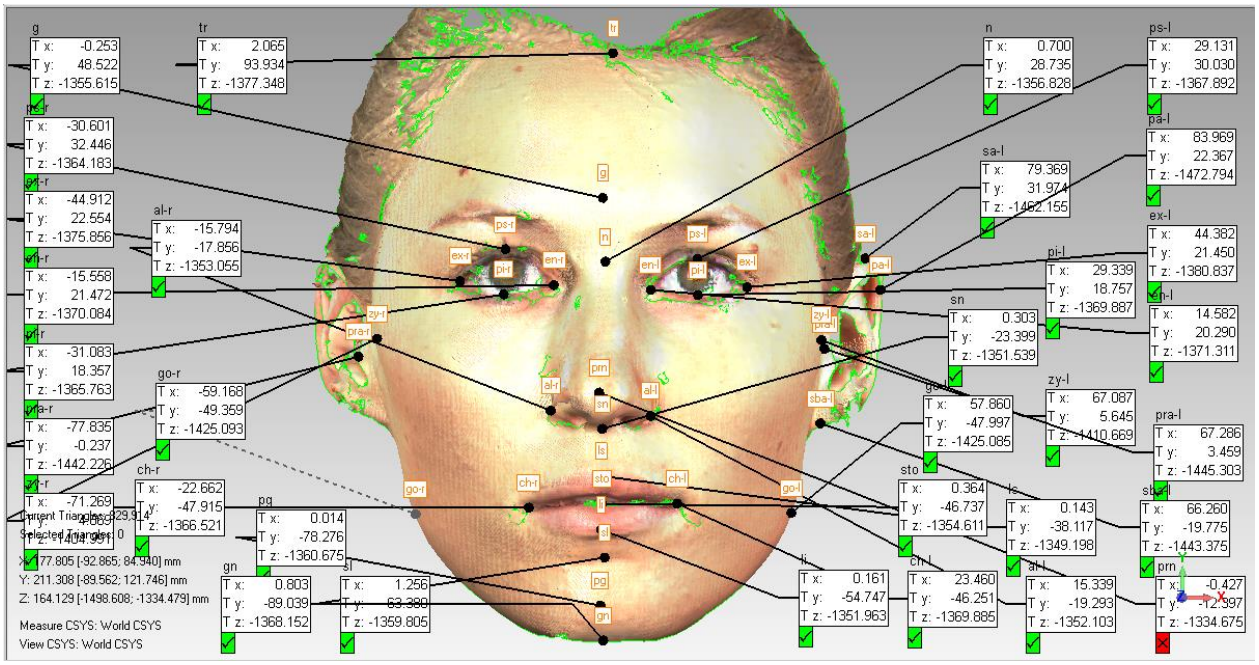


Figure 29. Illustration of 32 facial landmarks allocated on face with their corresponding coordinates.

The following landmarking protocol provides a list of 32 facial landmarks that were used in this study, and describes their anthropological location, according to Farkas et.al [17, 18] and Swennen et. al. [339]. The left (l) and right (r) sides of the relevant landmarks identified according to corresponding anatomical sides of the face. The names in parentheses are commonly used alternative anthropometric names of the corresponding landmark.

Table 4. Facial landmarks used in the project.

Landmark	Number	Abbreviation	Left (l) or Right (r)	Description	Notes
Gnathion (Menthon)	1	gn		The lowest anterior midpoint of the chin, located on the skin surface in front of the identical landmark of the mandible	Identical to the bony gnathion. The lowest point used in measuring facial height
Pogonion	2	pg		The most anterior midpoint of the chin, located on the surface in front of the identical bony landmark on the mandible	
Sublabiale	3	sl		Determines the lower border of the lower lip or the upper border of the chin	
Labiale inferius	4	li		The midpoint of the lower vermillion line	
Stomion	5	sto		The imaginary point at the crossing of the vertical facial midline and the horizontal labial fissure between gently closed lips, with teeth shut in the natural position	
Labiale superius	6	ls		The midpoint of the upper vermillion line	
Cheilion	7	ch(l)	Left	The point located at each labial commissure	Located on very comers of the mouth, not the lips
Cheilion	8	ch(r)	Right	The point located at each labial commissure	Located on very comers of the mouth, not the

					lips
Gonion	9	go(l)	Left	The most lateral point on the mandibular angle close to the bony gonion	If the angle is flat or if there is a rich soft-tissue cover, determination of this point is very difficult
Gonion	10	go(r)	Right	The most lateral point on the mandibular angle close to the bony gonion	If the angle is flat or if there is a rich soft-tissue cover, determination of this point is very difficult
Subnasale	11	sn		The midpoint of the angle at the columnella base where the lower border of the nasal septum and the surface of the upper lip meet.	The location is different to the bony subnasion. The landmark is identified in base view of the nose or from the side.
Pronasale	12	prn		The most protruded point of the apex nasi.	This point is difficult to determine if the nasal tip is flat. In the case of the bifid nose, the more protruding tip is chosen for prn
Alare	13	al(l)	Left	The most protruded point of each alar contour	
Alare	14	al(r)	Right	The most protruded point of each alar contour	
Nasion (soft tissue)	15	n		The deepest point on the nasal bridge	

Glabella	16	g		The most prominent midline point between the eyebrows	
Trichion	17	tr		The point of the hairline in the midline of the forehead	Cannot be determined on a balding head
Endocanthion	18	en(l)	Left	The point at the inner commissure of the eye fissure	
Exocanthion	19	ex(l)	Left	The point at the outer commissure of the eye fissure	
Palpebrale superius	20	ps(l)	Left	The highest point in the midportion of the free margin of each upper eyelid	
Palpebrale inferius	21	pi(l)	Left	The lowest point in the midportion of the free margin of each lower eyelid	
Endocanthion	22	en(r)	Right	The point at the inner commissure of the eye fissure	
Exocanthion	23	ex(r)	Right	The point at the outer commissure of the eye fissure	
Palpebrale superius	24	ps(r)	Right	The highest point in the midportion of the free margin of each upper eyelid	
Palpebrale inferius	25	pi(r)	Right	The lowest point in the midportion of the free margin of each lower eyelid	
Zygion	26	zy(r)	Right	The most lateral point of each zygomatic arch	Identical to the bony zygion of the malar bones. If the angle is flat or if there is

					a rich soft-tissue cover, determination of this point is difficult
Zygion	27	zy(l)	Left	The most lateral point of each zygomatic arch	Identical to the bony zygion of the malar bones. If the angle is flat or if there is a rich soft-tissue cover, determination of this point is difficult
Superaurale	28	sa(l)	Left	The highest point on the free margin of the ear lobe	
Subalare	29	sba(l)	Left	The lowest point on the free margin of the ear lobe	
Postaurale	30	pa(l)	Left	The most posterior point on the free margin of the ear	
Tragion	31	t(l)	Left	The notch on the upper margin of the tragus	
Tragion	32	t(r)	Right	The notch on the upper margin of the tragus	

All the facial landmarks were allocated manually, using 3D images generated by the Geomagic software.

2.9.2. Linear measurements

The data for 32 facial landmarks were processed in Excel to calculate each of the 54 linear measurements. The general formula for calculating a distance between two landmarks in the Euclidean space is: $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$ where 'x, y and z' are the coordinates of each landmark in the Cartesian space. The formula was used in the Excel spreadsheet for automatic calculation of the linear distances.

For images of a good quality (full set of 32 facial landmarks available), a full set of measurements was generated. For images of a poor quality, a partial set of measurements was recorded.

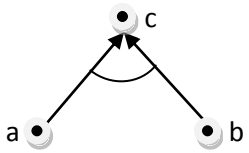
The data for the following facial measurements were generated for each 3D image:

- Total face height: tr-gn
- Face width: zy-zy
- Morphological face height: n-gn
- Physiognomical face height: n-sto
- Lower profile height: prn-gn
- Lower face height: sn-gn
- Lower third face depth: t(l)-gn
- Middle face depth: t(l)-prn
- Middle face height (right): go(r)-zy(r)
- Middle face height (left): go(l)-zy(l)
- Middle face width 1: t(r)-t(l)
- Middle face width 2 (left): zy(l)-al(l)
- Middle face width 2 (right): zy(r)-al(r)
- Upper face depth: (left): t(l)-tr
- Upper face depth: (right): t(r)-tr
- Upper third face depth: t(l)-n
- Forehead height: g-tr
- Extended forehead height: tr-n
- Glabella –Gnathion distance: g-gn
- Supraorbital depth: t(l)-g
- Trichion – Zygion distance (left): tr-zy(l)
- Trichion – Zygion distance (right): tr-zy(r)
- Nasion - Zygion distance (left): n-zy(l)

- Nasion - Zygion distance (right): n-zy(r)
- Zygion – Gnathion distance (left): zy(l)-gn
- Zygion – Gnathion distance (right): zy(r)-gn
- Intercanthal width: en-en
- Biocular width: ex-ex
- Eye fissure width (left): en(l)-ex(l)
- Eye fissure width (right): en(r)-ex(r)
- Eye fissure height (left): ps(l)-pi(l)
- Eye fissure height (right): ps(r)-pi(r)
- Ear height (left): sa(l)-sba(l)
- Ear width (left): t(l)-pa(l)
- Nasal bridge width: n-prn
- Nose height: n-sn
- Nose width: al-al
- Nasal tip protrusion: sn-prn
- Ala length (left): prn-al(l)
- Ala length (right): prn-al(r)
- Gonion - Trichion distance (left): go(l)-tr
- Gonion - Trichion distance (right): go(r)-tr
- Gonion – Glabella distance: g-pg
- Pronasale - Gonion distance (left): prn-go(l)
- Pronasale - Gonion distance (right): prn-go(r)
- Chin height: sl-gn
- Mandibular region depth (right): t(r)-gn
- Mandible width: go-go
- Mandible height: sto-gn
- Lower jaw depth (left): gn-go(l)
- Lower jaw depth (right): gn-go(r)
- Mouth width: ch-ch
- Upper vermilion height: ls-sto
- Lower vermilion height: li-sto

2.9.3. Angular measurements

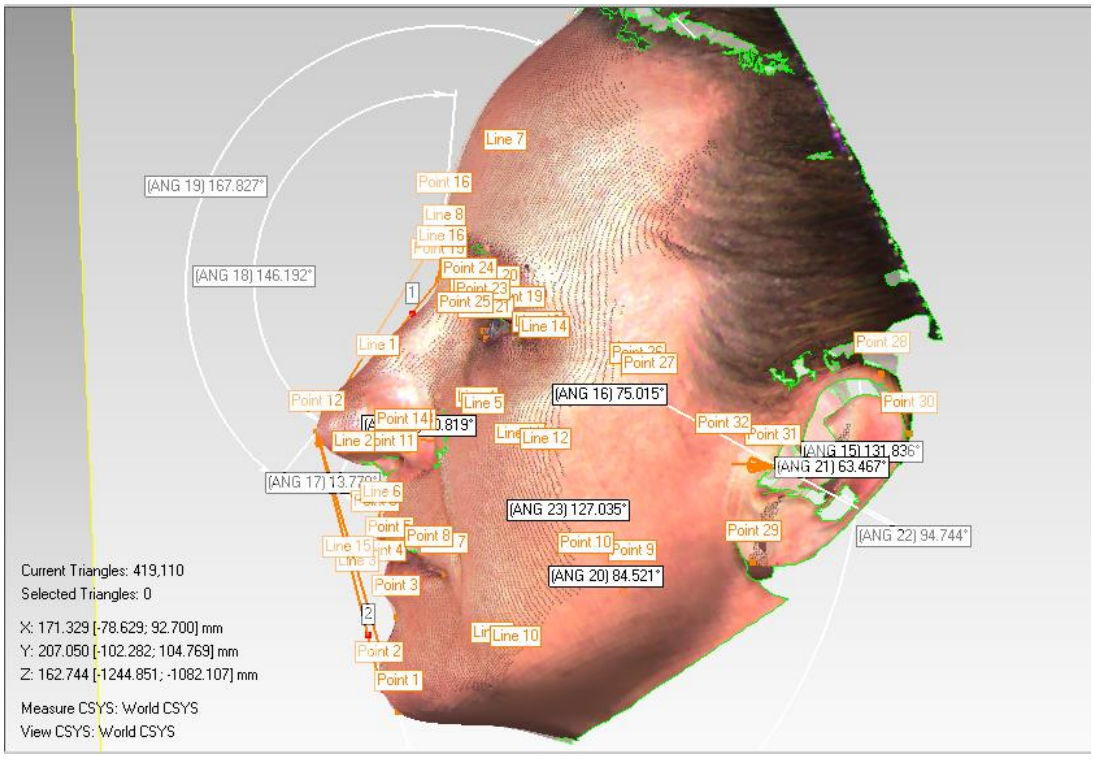
Ten angular measurements were calculated based on previously allocated Euclidean coordinates for facial landmarks. Angular measurements were calculated between two 3D vectors of the three corresponding landmarks. As shown at the following scheme, each three landmarks for any angular measurement form two vectors ("ac" and "bc"), and the angle is measured between vectors:



The general formula for calculating an angle between two vectors in the Cartesian coordinates system is: $\cos\theta = (a.b)/(|a||b|)$

Figure 30 demonstrates all the angular distances, formed by corresponding vectors on a facial image with (a) and without (b) removing the facial background.

a)



b)

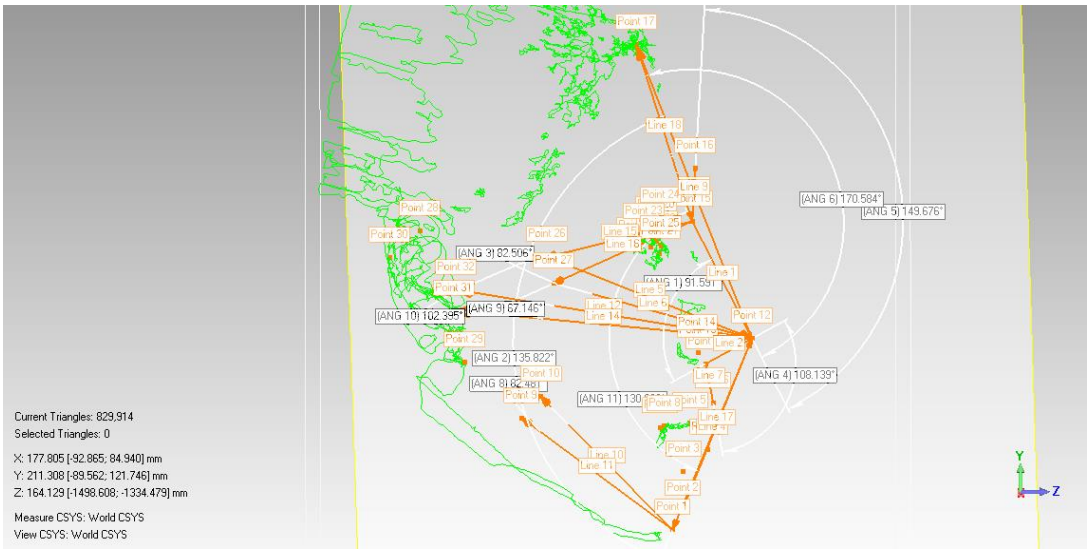


Figure 30. An example of angular distances on a 3D image a) with and b) without facial surface texture.

The following list summarizes ten angular measurements, obtained from 3D facial images.

- Nasal tip angle (n-prn-sn)
- Nasal vertical prominence angle (tr-prn-gn)
- Transverse nasal prominence 1 angle (zy(l)-prn-zy(r))
- Transverse nasal prominence 2 angle (t(l)-prn-t(r))
- Nasolabial angle (prn-sn-ls)
- Nasofrontal angle (g-n-prn)
- Nasion depth angle (zy(l)-n-zy(r))
- Nasomental angle (n-prn-pg)
- Forehead nasal angle (tr-n-prn)
- Chin prominence angle (go(l)-gn-go(r))

2.9.4. Ratios

The data acquired from linear measurements were used to calculate 21 facial indices (ratios), according to the following list:

- Forehead height ratio (tr-n)x100/(go(r)-go(l))
- Upper face height ratio (n-sn)x100/(go(r)-go(l))
- Lower face height ratio: (sn-gnx100/(go-go)
- Anterior face height 1 ratio: (n-gn)x100/(go-go)
- Anterior face height 2 ratio: (n-gn)x100/(zy-zy)
- Face height index: (n-gn)x100/(tr-gn)
- Upper – Lower face ratio: (tr-g)x100/(sn-gn)
- Upper face height ratio: (n-sn)x100/(sn-gn)
- Upper face width ratio: (n-sn)x100/(zy-zy)
- Total anterior face height ratio: (tr-gn)x100/(zy-zy)
- Mouth width ratio: (ch-ch)x100/(en-en)
- Mandible – Face width ratio: (go-go)x100/(zy-zy)
- Mandible index: (sto-gn)x100/(go-go)
- Mandible – Interexocanthion distance ratio (go-go)x100/(ex-ex)
- Interendocanthion distance ratio: (en-en)x100/(al-al)
- Intercanthal index: (en(R)-en(L))x100/(ex(R)-ex(L))
- Intercanthal – Intracanthal index: (ex(R)-en(R))x100/(en(L)-ex(L))
- Nasal index: (al-al)x100/(n-sn)
- Nose-face height index: (n-sn)x100/(n-gn)
- Nose-face width index: (al-al)x100/(zy-zy)
- Nasal tip protrusion – nose width index: (sn-prn)/(al-al)
- Nasal tip protrusion –Nose height index: (sn-prn)x100/(n-sn)

2.9.5. Principal components

All the craniofacial measurements and indexes (n=85) generated in this study were used for calculation of 20 principal components, as discussed in details in Section 4.6.

2.10. Genotyping methods

A set of 587 DNA samples was genotyped by sequencing. DNA samples were processed according to the following procedures:

2.10.1. Custom Ampliseq protocol for library preparation

A custom Ampliseq library was prepared from 10 ng of genomic DNA which was amplified using an IonAmpliseq library kit in conjunction with a custom Ampliseq primer pool, according to manufacturer recommendations [340]. The original amplification reaction was split into two 10µl reactions for each of the multiplex pools, as per manufacturer's recommendations. The master mix included 2µl of 5x Ion AmpliSeq™ HiFi Master Mix and 5µl of each 2x Ion AmpliSeq™ Primer Pool. Each primer pool consisted of approximately 840 primer pairs. Each library was prepared for up to 32 DNA samples in a 96 well plate format. The amplification conditions included an initial 'hold' step of 99^0C for 2 minutes, 15 cycles of 99^0C for 15 seconds followed by 60^0C for 8 minutes and a final hold step of 10^0C for up to 1 hour.

Following PCR, two pools of amplicons for each sample were mixed in a final volume of 20 µl. The resulting amplicons were treated with 2µl FuPa reagent to partially digest the primers and phosphorylate the amplicons. The sample plate was placed in a thermal cycler and incubated at 50^0C for 10 minutes, followed by 10 minutes incubation at 55^0C and then by 60^0C incubation for 20 minutes.

Subsequent to primer digestion, the amplicons were ligated to Ion Adapters with up to 32 unique barcodes (according to the number of samples). The master mix consisted of 4µl Switch solution, 2µl of Ion AmpliSeq barcodes and 2µl of DNA ligase for each well. The ligation reaction was performed at 22^0C for 30 minutes and then at 72^0C for 10 minutes. The ligated library products were purified using the Agencourt AMPure^{TM} XP reagent according to manufacturer recommendations [341].

The final libraries were quantified using an Ion Library Quantitation Kit as per manufacturer recommendations. The quantitative Real Time PCR mix was prepared in a 96-well format, either manually or using an epMotion 5075 robot (Eppendorf South Pacific, North Ryde, NSW). The amplification reaction was performed in 10µl (half of the recommended reaction volume). The master mix consisted of 5µl of 2x Ion TaqMan master mix and 0.5µl 20x Ion TaqMan Assay. The volume of DNA sample was 4.5µl

per reaction. An *E.coli* DH10B Ion Control Library was used as a standard. The amplification step included 50^0C for 2 minutes, 95^0C for 20 seconds, followed by 40 cycles of 95^0C for 3 seconds and 60^0C for 30 seconds each. The quantitative PCR was performed using a CFX96 Real-Time system (BioRad), as per manufacturer recommendations. The quality of the amplified libraries (e.g. amplicon sizes) was not confirmed as recommended due to unavailability of the Bioanalyzer instrument (Agilent).

The pre – PCR and post – PCR steps of the library preparation were performed in two separate dedicated hoods. Each step of the protocol was followed by a decontamination step, using 10% hypochloride solution and UV irradiation. Figure 31 summarises the Ampliseq library preparation step:

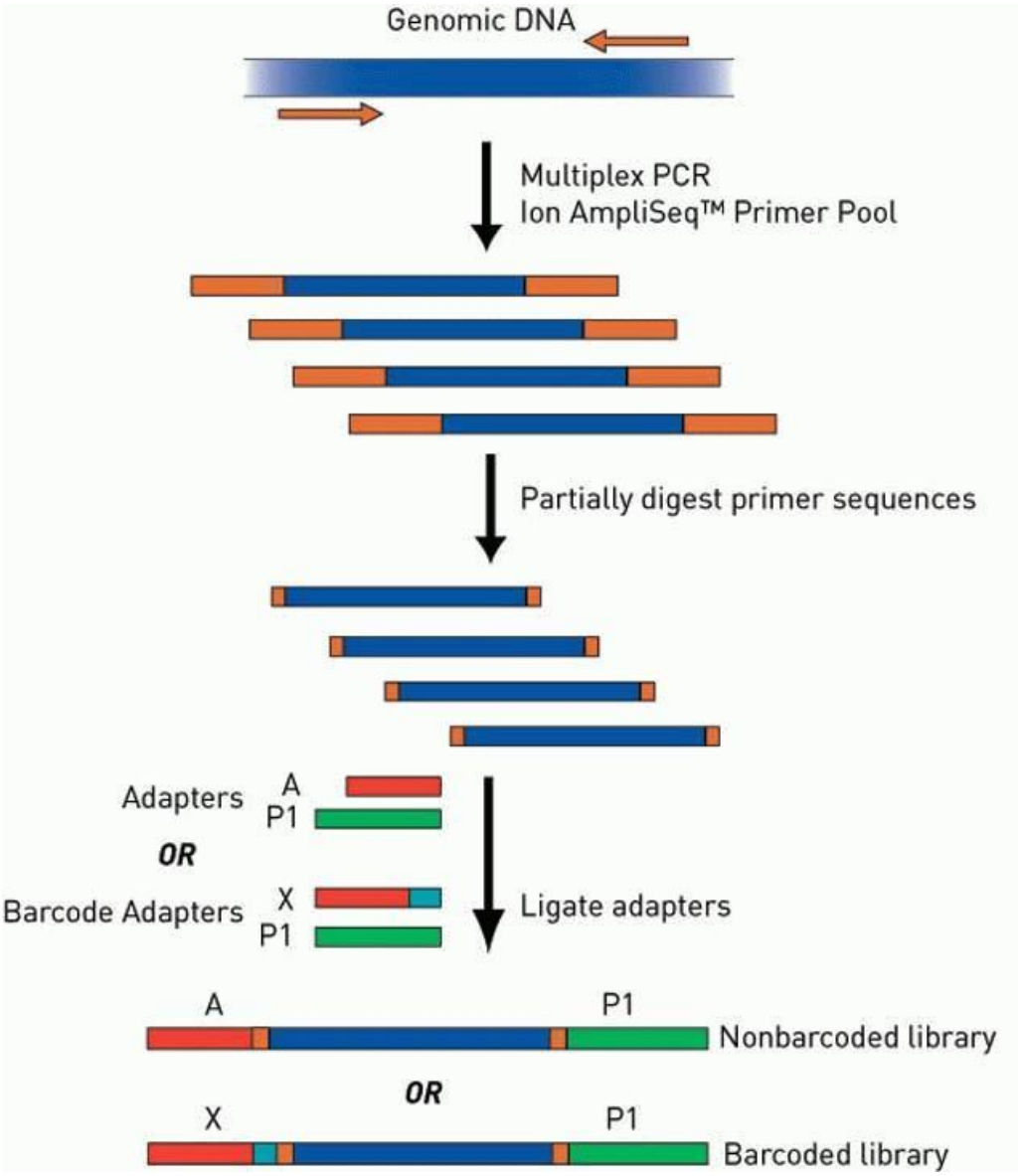


Figure 31. Ampliseq Library preparation summary.

2.10.2. Template preparation

Each library was diluted to approximately 10 pM to 20 pM and mixed in equimolar ratios. Template-positive Ion PGM™ Template 200 Ion Sphere™ Particles (ISPs) with 200 base-pair average insert libraries were automatically prepared using either OneTouch™, OneTouch™ DL or OneTouch™ 2 instruments. The ISPs were subsequently enriched using a semi-automated protocol on the Ion OneTouch™ ES instrument. All steps were performed according to manufacturer's recommendations [342].

2.10.3. Sequencing

The sequencing step was performed on the Personal Genome Machine (PGM) according to manufacturer recommendations [343]. Following the cleaning and initialization steps, the 316 chip was loaded with enriched ISPs (following template preparation step) and sequenced on the PGM. The sequencing round lasted approximately 8 hours and included sequencing of two 316 chips (up to 32 libraries per chip).

2.10.4. Data analysis

During the sequencing process, the initial unaligned sequence data were automatically transferred to the Torrent Suite Server (TSS). The TSS automatically performed sequence alignment according to the reference sequence, quality control and data analysis, such as basic variant calling. The initial data analysis process is summarised in Figure 32.

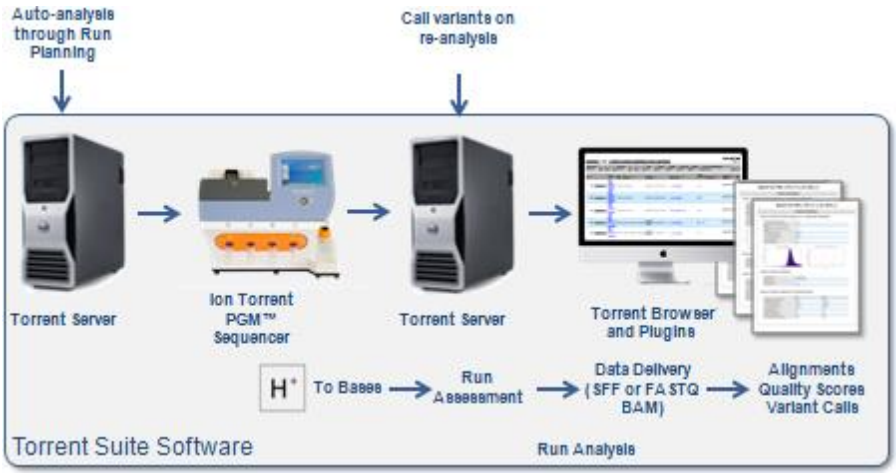


Figure 32. Illustration of the data analysis workflow.

At the end of each sequencing run, the TSS software generated a summary, with the following quality control parameters of the sequencing run performance:

- **ISP loading density**. This parameter represents the percentage of chip's wells loaded with ISPs. The red-coloured areas in Figure 33 represent fully loaded wells, while the yellow colour represents less loaded and blue and green areas shows very poorly loaded wells. The blue areas usually represent air bubbles, which can be "trapped" inside the chip, during the loading process. Alternatively, these areas may be caused by a technical failure in the chip manufacturing process.

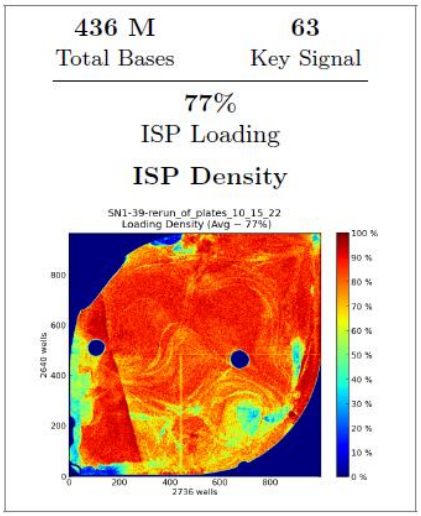
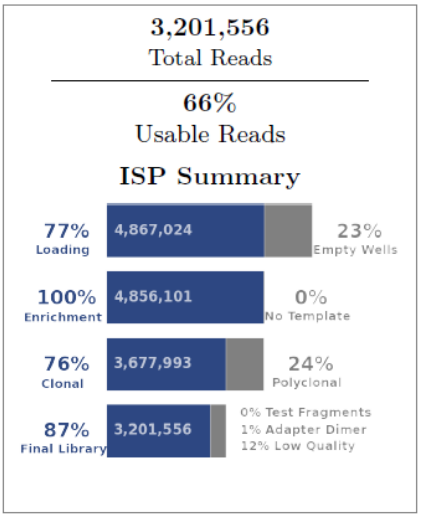


Figure 33. Loaded chip image, illustrating loading density. In this image, red indicates good loading, yellow is passable, and green and blue (air bubbles) show very poor loading.

- **Usable reads**, representing various parameters of the ISP beads performance (Figure 34), such as:

- The percentage of loaded vs. empty chip wells. This value depends on various parameters, such as the number of beads available and a physical distribution of ISPs during the loading process.
- The percentage of enriched ISP. This value depends on the initial concentration of the libraries, used for enrichment and efficiency of the template preparation on OneTouch™ instrument.
- The percentage of live ISPs. This value represents a number of wells, which contained an ISP with a signal of sufficient strength to be associated with the library fragment of test fragment key.
- The test fragment key. This parameter represents a number of live ISPs that were identical to the test fragment key signal.
- The percentage of polyclonal beads. This value represents the number of beads that have more than a single targeted clone and as a result, cannot be used for sequencing and alignment. Based on manufacturer recommendations, a sequencing run with polyclonality values of less than 40% is considered of a good quality.
- The percentage of adapter-dimer. This value refers to the formation of a sequencing adaptor-dimer, with no targeted insert or a very short insert. The fraction of such reads is typically less than 1%.
- The percentage of low quality reads. Any inserts of less than 8 base pairs or low-quality base calls at the 3’ ends are automatically trimmed and removed from the final output.



Addressable Wells	6,348,245	
With ISPs	4,867,024	76.7%
Live	4,856,101	99.8%
Test Fragment	15,964	00.3%
Library	4,840,137	99.7%

Library ISPs	4,840,137	
Filtered: Polyclonal	1,178,108	24.3%
Filtered: Low Quality	415,505	08.6%
Filtered: Primer Dimer	44,968	00.9%
Final Library ISPs	**3,201,556**	**66.1%**

Figure 34. An example of sequencing run statistics.

- **Read length** represents the distribution of sequenced fragments in the run. The ideal distribution of fragments in the custom Ampliseq assay used in this study was aimed to be shifted towards the 150-200bp region (Figure 35).

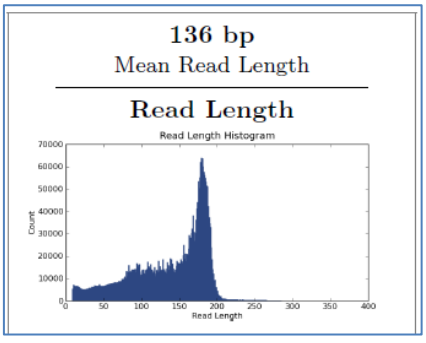


Figure 35. An illustration of amplicon length distribution in a single sequencing run.

- **Run quality following alignment (AQ17/ AQ20)**. This value represents the greatest position in the read at which the accuracy in the bases meets the accuracy threshold (see an example in Figure 36). The AQ17 length of a read is the greatest length at which the read error rate is 2% or less and the AQ20 length is the greatest length at which the error rate is 1% or less.

	AQ17	AQ20	Perfect
Total Number of Bases [Mbp]	393 M	333 M	258 M
Mean Length [bp]	134	118	94
Longest Alignment [bp]	314	307	303
Mean Coverage Depth	0.1	0.1	0.1

Figure 36. An illustration of run quality metrics.

The Ion Torrent data analysis process includes several continuous steps and is summarised in Figure 37.

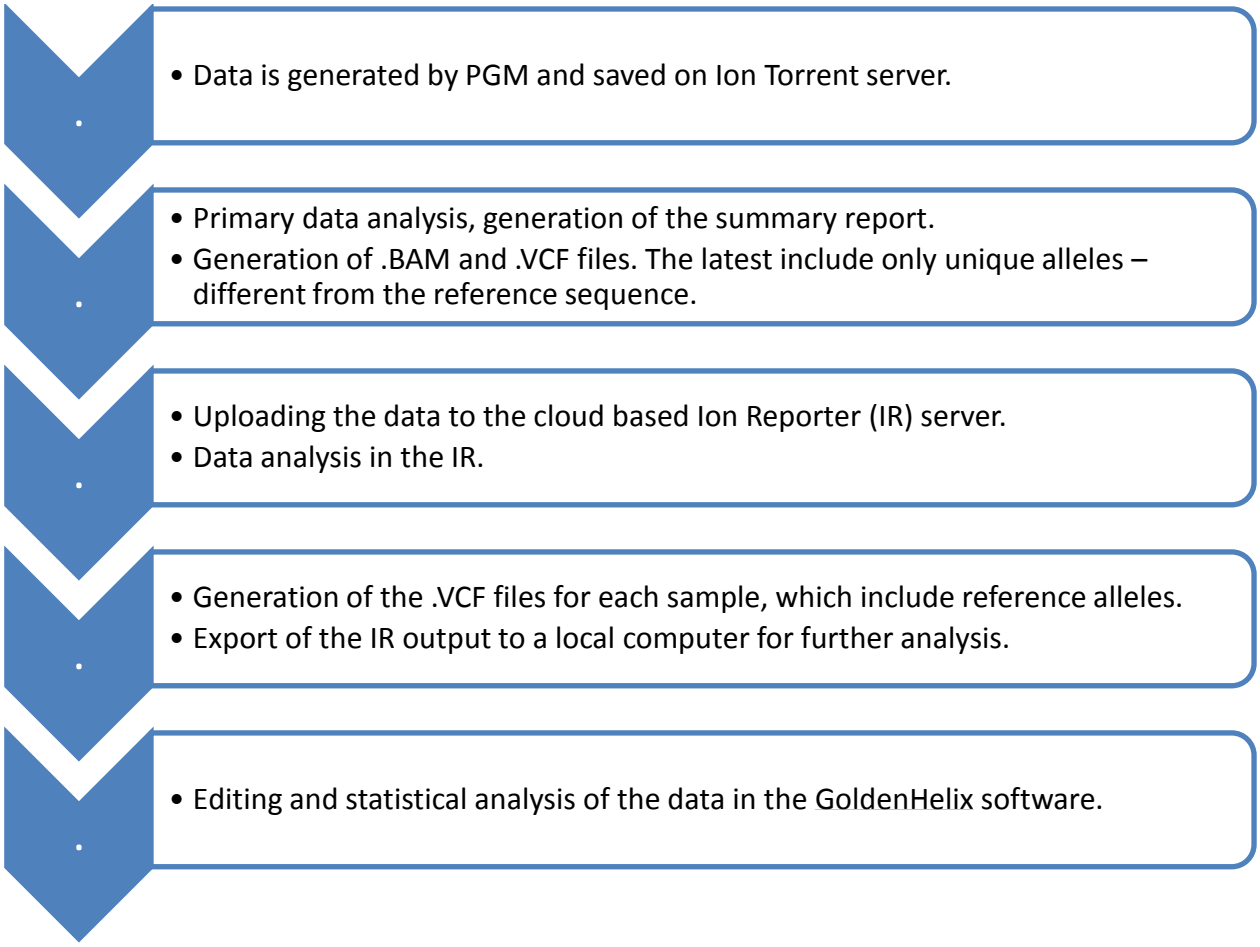


Figure 37. A summary of the Ion Torrent data analysis workflow.

2.10.5. Variants annotation using the Ion Reporter software

The aligned sequences in BAM format were uploaded to the Ion Reporter (IR) cloud-based software for further analysis. Ion Reporter software provided the opportunity to annotate and export all the variants, including the reference calls, which was not possible with the build-in Variant Caller in the Ion Torrent Suite software. The IR output was exported in the VCF format, which included allele calls and various sequencing statistics (such as sequencing depth and quality, allele frequency and markers type and chromosomal location) for each sample.

2.11. Statistical analysis

2.11.1. Data dimensionality reduction

In order to more efficiently manage the relatively large number of craniofacial measurements (n=92), a Principal Component Analysis (PCA) was conducted on the craniofacial measurements dataset, using the PASW Statistics 18 release version (SPSS, Inc., Chicago, IL.). A more detailed explanation of this method is provided in Section 4.6.

2.11.2. Markers association analysis

Ancestry association analyses were performed using SNP & Variation Suite v7 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) and replicated using PLINK v1.07 software [344]. A stepwise linear regression analysis with cofactors such as sex and ancestry were implemented to find potential associations of specific phenotypic traits with genetic markers. A more detailed explanation of this method is provided in Chapters 3 and 5.

2.11.3. Prediction analysis of phenotypic traits and ancestry

Numerous craniofacial and pigmentation phenotypic traits and ancestry were predicted using stepwise logistic regression models. Multiple determination coefficients (R^2) of these models were used to estimate the predictive power of the model. A more detailed explanation of this method is provided in Chapter 5.

Chapter 3

Assessment of samples collection, DNA extraction and genotyping equipment optimisation

3.1. General Introduction

This chapter refers to aims 1-3 of this project, which were:

AIM 1: To review the literature and relevant web resources in order to create a list of candidate genes and SNPs, potentially involved in normal craniofacial appearance. To generate a shortlist of SNPs, targeted for subsequent genotyping by sequencing.

AIM 2: To collect a database of 300-500 DNA samples, along with 3D facial images and relevant phenotypic information. To analyse the 3D images for a set of craniofacial landmarks and various facial measurements.

AIM 3: To validate and optimize the various methods used in this project, such as DNA extraction, genotyping and 3D image analysis.

Specifically, this chapter focuses on the results of following studies:

- DNA sample collection.
- Validation and optimisation of the DNA extraction methods.
- Candidate genes and SNP selection.
- Evaluation of SNP genotyping platforms.

Prior to performing genotyping all relevant methods and use of equipment were optimised, when needed. This was necessary for the subsequent association and prediction studies, which were the main aims of this project.

3.2. Sample collection

3.2.1. Introduction

The sample collection process occurred over a period of three and a half years. In order to make the sample collection procedure more convenient and efficient, an online booking web site, which provided a user-friendly interface and the opportunity for volunteers to book an appointment for the 3D scanning and DNA collection was

created. The use of the online schedule helped with the logistics of the sample collection process. The booking web site is accessible through the following link:

http://www.supersaas.com/schedule/DNA_Database/Craniofacial_project

In order to recruit volunteers, the research project was advertised using flyers and media releases, as well as via personal communication.

3.2.2. Methods

DNA samples and 3D images were collected as detailed in the Chapter 2.

3.2.3. Results and Discussion

The sample collection procedure resulted in recruiting 641 samples, with 107 individuals providing a DNA sample only, without the 3D image. However, of these 107, fourteen (14) DNA samples were used for pigmentation and ancestry association studies only (not for the craniofacial traits association study). As a result, there were 534 volunteers who donated both a DNA sample and a 3D facial image. In addition, 39 individuals were sourced from the existing database, collected between 2000 to 2009. These samples were accompanied with pigmentation and ancestry information. Overall, 587 DNA samples were used for genotyping purposes. The mean age of the 587 volunteers was 27.4 year old with standard deviation of 10 years.

The volunteers were of various ancestries, with the majority (70%) being Caucasian. A sample set comprised of different ancestries can potentially introduce a bias in statistical analysis, due to the difference in allele distribution in various populations. However, this issue was addressed by performing a statistical correction to address the population stratification issue in the association analysis (as detailed in Chapter 5). Table 5 summarises the volunteer statistics on sex and ancestry.

Table 5. Samples collected in the current study as categorised by sex and ancestry.

	Trait	No. samples	Percentage
Sex	Male	231	39.5
	Female	354	60.5
	Total	585	100
Ancestry	Aboriginal	3	0.5
	African	23	3.9
	Admixture	45	7.7
	Asian	55	9.4
	Caucasian	409	69.9
	Indian	46	7.9
	Other	4	0.7
	Total	585	100

Generally, the bigger the sample size, the greater the statistical significance of the results and the lower the false positive rate. However, collecting and genotyping of larger sample numbers is not always feasible, since it is more time-consuming and costly. Furthermore, very large sample sets often involve collaborative efforts of several researchers, which is also not always possible. The final number of samples used in this study may still be considered small for an association study, while the use of a targeted gene/SNP approach mitigates this.

An effective sample size is defined as the minimum number of samples that achieves adequate statistical power. A statistical power of 80% is considered sufficient for large-scale association studies [345, 346]. However, testing a large number of SNP markers leads to a large number of multiple comparisons and thus increases false positive rates. In order to correct the false positive errors (type I error), either the Bonferroni or the false discovery rate correction is generally applied [347]. Given that the more often used (and more stringent) Bonferroni correction involves dividing the p-value significance threshold by the number of markers tested, a large number of samples is generally needed to compensate for this correction. As a result, most GWAS which analyse association of millions of markers with a specific trait, require thousands of samples to be collected. It has been shown that testing a single SNP in a disease association study would require 248 cases, while testing 500,000 markers would require at least 1,206 cases [348]. Taking into account the relatively small number of markers tested in this project (approximately 3,000 SNPs) due to the candidate markers targeted approach, the

sample size used in this study (n=587) can be considered sufficient. In fact, results of the association analysis demonstrated that the sample size used was statistically effective (Chapter 5). However, as in most association studies, an additional set of samples might be required in order to replicate these results.

3.3. DNA extraction methods evaluation

3.3.1. Introduction

Different DNA extraction methods were evaluated in order to find a fast, easy, robust and cheap method as an alternative to the automated, but costly DNA extraction procedure (EZ1 extraction). Given the high number of buccal swabs processed (four per individual), a cheap yet effective DNA extraction method was required to meet the budget frame of this project.

3.3.2. Methods

Four DNA extraction methods were evaluated:

- Qiagen EZ1 purification kit (n=68)

  This protocol utilises a semi-automated DNA purification by a BioRobot EZ1 workstation purifying up to 6 samples per one extraction. Cost per sample: approximately AU $11.
- Promega DNA IQ extraction kit (n=24)

  This is a manual protocol that uses magnetic particles, specially designed for forensic purposes, which are able to capture and purify DNA molecules. There is no sample number limitation per extraction, although usually up to 24 samples are processed for easier handling. Cost per sample: approximately AU $6.5.
- Qiagen miniprep kit (n=24)

  This method uses various lysis and precipitation buffers to manually purify up to 20 µg of DNA per sample from a variety of tissues. There is no sample number limitation per extraction, although usually up to 24 samples are processed for easier handling. Cost per sample: approximately AU $3.5.
- Isohelix purification kit (n=70)

  This is a manual protocol. It is specifically formulated to produce high DNA yield and purity from buccal samples. There is no specific sample number limitation per extraction, although usually up to 24 samples are processed for easier handling. Cost per sample, including buccal swab: approximately AU $3*.

* The price for all purification kits, except Isohelix, did not include swabs cost.

One set of DNA samples (n=24) was used for the comparison between Qiagen miniprep and Promega DNA IQ kits, while another set of samples (n=68) was used for comparison between Qiagen EZ1purification procedure and Isohelix kit. Two different sets of duplicates were used due to the limited number of buccal swabs collected per person (n=4), while each two (2) were used for a single extraction. The DNA extractions using four (4) kits were performed over a period of three (3) months.

3.3.3. Results and Discussion

The results show that both the DNA IQ and Miniprep kits generated significantly lower DNA yields, compared to the EZ1 and Isohelix extractions. The Isohelix method resulted in the highest yield, although there was greater variability between the samples (Table 6). However, the Isohelix method was cheaper, simpler and provided higher throughput. On the other hand, the purity of the DNA extracted with Isohelix kit was lower, due to presence of the cellular debris (data not shown). Nevertheless, this fact did not compromise the subsequent MPS output.

Based on these results and extraction cost, Isohelix swabs and extraction kits were chosen for sample collection and DNA extraction in this project.

Table 6. Real Time PCR quantification results of DNA extraction from buccal swabs, using 4 different extraction protocols. The DNA concentration is shown in µg/µl and normalised to 100 µl final volume. Note that samples used in EZ1 vs. Isohelix and DNA IQ vs. Mini prep extraction procedures respectively, are non-identical.

Concentration (ng/µl)				
	EZ1 extraction	**Isohelix extraction**	**DNA IQ**	**Mini prep**
Median	30.86	41.83	4.41	13.44
SD	11.61	21.01	6.3	18.6

3.4. Candidate genes and markers selection

3.4.1. Introduction

The main aim of this research was to find SNPs which affect normal craniofacial variation. A candidate gene approach was chosen as the most appropriate and feasible method for this project (as discussed in Chapter 1). The process of candidate gene and SNP selection involved literature review and screening various bioinformatical web resources. This approach provided information on genes and respective polymorphisms, potentially influencing craniofacial morphology (discussed in details in Chapters 1 and 2).

The number of markers planned to be genotyped initially was 96 SNPs, due to limitation of the Golden Gate assay, as detailed in Section 3.5. However, with the subsequent opportunity of the greater multiplexing capability of the Ion Torrent platform, the original craniofacial SNPs number was increased and also extended to markers, which have been shown to be associated with pigmentation traits, ancestry as well as identity informative SNPs, INDELs and STRs markers. This assay is expected to provide a maximum information from the DNA sample, using a single workflow that consumes less amount of a limited template.

This section provides a summary of the candidate genes and markers selection process and details of their annotation analysis.

3.4.2. Methods

Two main complementary strategies were used to generate a list of candidate markers. The first focused on searching the literature for candidate genes, involved either in normal craniofacial variation or in craniofacial malformations in both humans and model organisms. The resulting set of SNPs and genes was further screened for high Fst SNPs (greater than 0.25) as well as potentially functional polymorphisms, such as non-synonymous SNPs, markers in transcription factors binding sites and splicing sites, using web resources listed in Section 2.5. The second approach focused on searching for genes with high Fst SNPs and dedicated databases of ancestry informative

markers (AIMs). The resulting set of SNPs was then screened for genes potentially involved in the craniofacial embryogenesis.

The web resources used for searching candidate genes and markers are listed in Section 2.5.

3.4.3. Results and Discussion

Both searching methodological approaches were performed simultaneously and pinpointed 1,088 candidate genes and intergenic regions. However, 592 of these genes showed no clear link with craniofacial development and were therefore not included. Most of these 592 genes were originally selected based on high Fst values of respective markers. This process resulted in 496 genes and intergenic regions potentially involved in regulating normal craniofacial morphology. These 496 genes/regions were screened for non-synonymous and potentially functional SNPs, which resulted in 269 genes and intergenic regions after removing genes that do not possess such markers, as well as SNPs with no evidence of high population differentiation ($Fst \leq 0.25$). Subsequent analysis of these 269 genes/regions using various resources for functional annotation, resulted in 137 candidate genes/regions, possessing 10,746 markers. Notably, the majority of these SNPs are located in introns and intergenic regions (as discussed in Chapter 5). This list was further analysed for tag SNPs, in order to reduce the number of markers for genotyping and subsequent analysis, based on the genotyping platform that was available (detailed in section 3.5). Following the availability of a MPS platform to this project, forty (40) additional genes were selected from recently published articles, all focusing on normal variation in craniofacial morphology, as discussed in Sections 1.4.1 and 1.4.3.

The final list comprised 177 candidate genes and intergenic regions, which were further screened for high Fst markers, non-synonymous SNPs, SNPs in transcription binding sites and SNPs in splicing sites, as summarized in Figure 38. The final screen revealed approximately 1,200 SNPs in the candidate genetic regions, potentially involved in determining craniofacial morphology and was used for initial primer design.

The final SNP panel was extended by adding 331 markers, shown to be associated with normal pigmentation, such as eye, skin and hair colour, 208 ancestry informative markers, 91 identity informative SNPs, 46 INDELs, 17 autosomal STRs, 15 Y-

chromosome STRs, 37 Y-chromosome SNPs, and 57 Mitochondrial SNPs and resulted in 1,985 targets, submitted to Life Technologies™ for initial primer design (summarized in Figure 39).

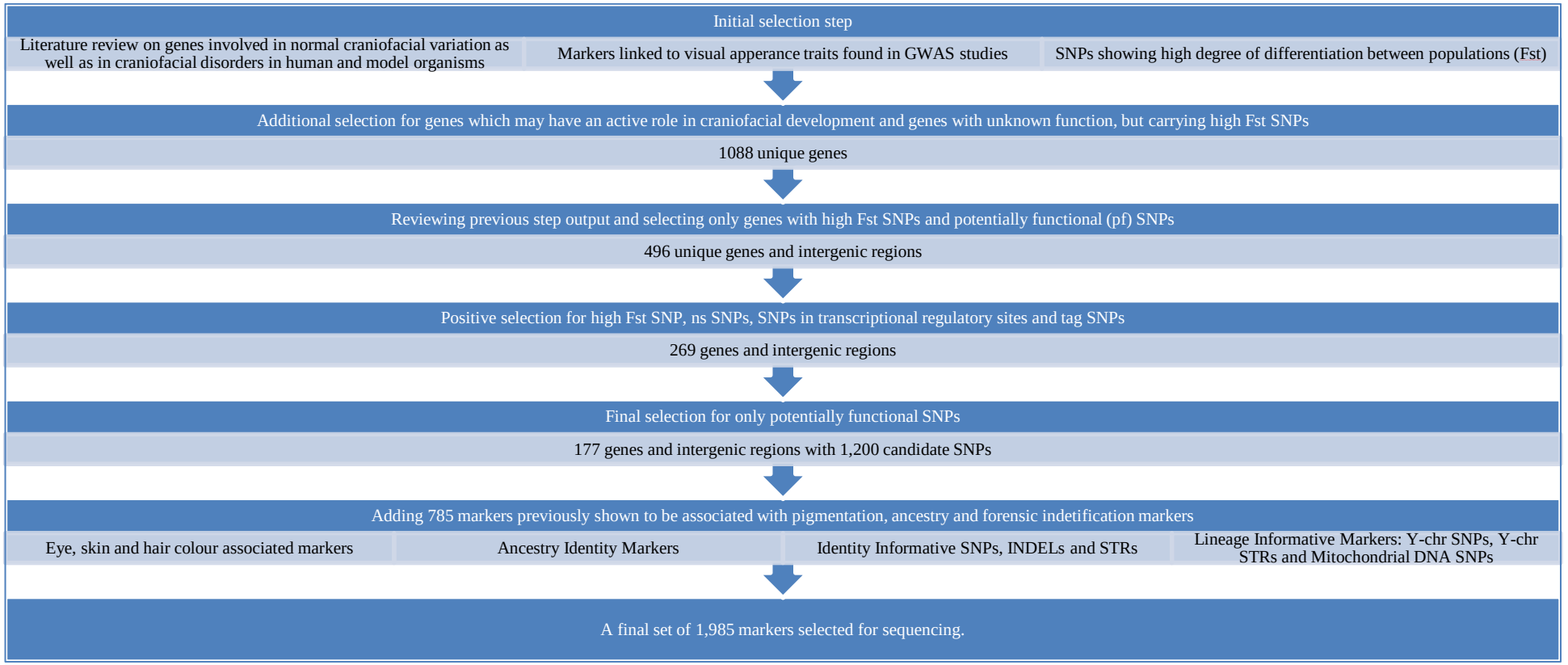


Figure 38. A summary of candidate genes and SNPs selection process.

All Y-chromosome STRs, all mitochondrial SNPs, 10 autosomal STRs and 522 craniofacial candidate SNPs failed the initial primer design process (approximately 20% of all markers). Although Life Technologies™ does not provide details for the technical issues with primer design failure, the most reasonable explanation is that these markers are located in highly polymorphic or repetitive regions of the genome (e.g. homopolymer repeats, Alu, LINE etc.). These genomic regions are known to be more unfavourable for the primer design process.

Additional markers that had been found to be in high linkage disequilibrium (LD) with the primary targeted SNPs that previously failed the primer design pipeline, were identified and used as substitutes. There were five iterations of SNP selection and primer design process, before the final set of SNP primers were made. The final list covered 1,933 primary targeted markers (Figure 39).

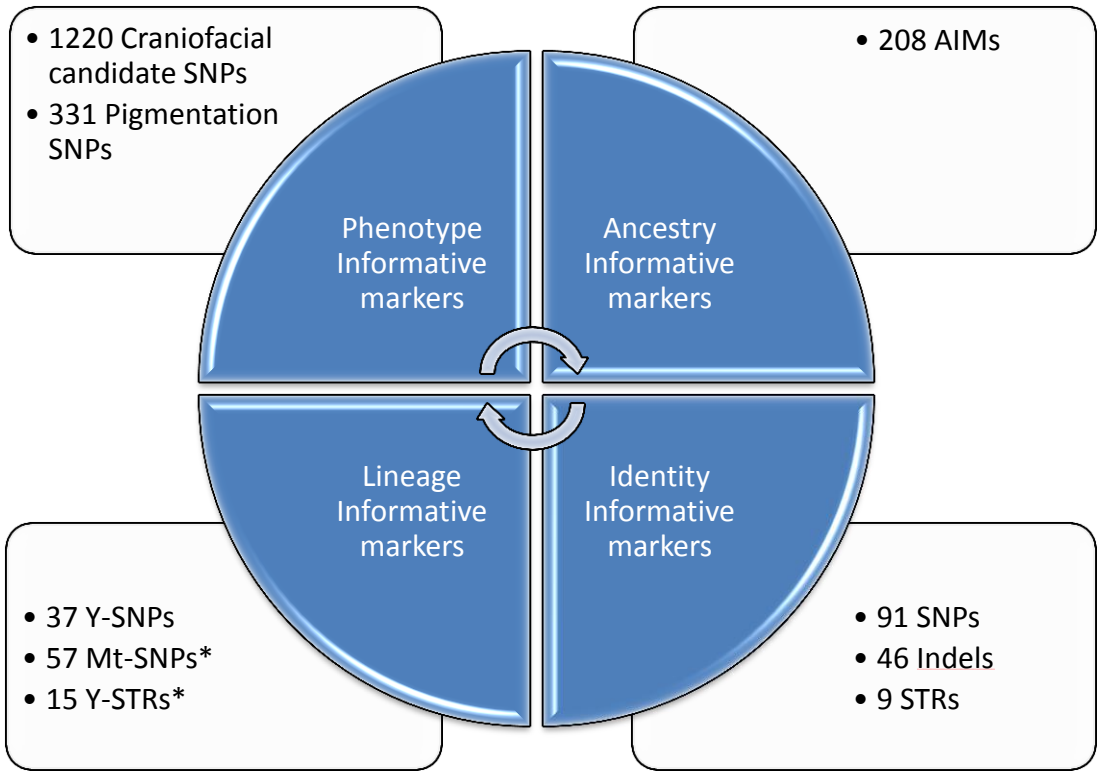


Figure 39. Initial SNP panel submitted to Life Technologies for primer design. *) These markers were removed from the final panel, due to primer design failure.

This list was transformed into a hotspot file in .BED format for the subsequent sequencing process. Given that each amplicon covers between 125 bp to 225 bp, each

amplicon could potentially cover more than one polymorphism. As a result, the initial amplicon list was screened for additional SNPs with 1% or less Minor Allele Frequency (MAF), using the UCSC genome browser.

The final Ampliseq panel size was 15.78 kb and covered 99.3% of the originally submitted targets, following the last design step. The final hotspot file of all the markers included approximately 6,500 known markers in 985 genes/regions and was used as a reference hotspot file for variant caller in the Ion Torrent Suite and Ion Reporter software. However, it should be noted that more than 50% of these markers are not common (less than 2% frequency) and as a result were filtered from the association analysis (Chapter 5).

The craniofacial candidate marker list (1,220 SNPs) was analysed using the Genomic Regions Enrichment of Annotations Tool (GREAT) web-based platform to visualize the genomic context of the amplicons covering targeted SNPs. This tool helps annotate non-coding genomic regions, which typically lack genetic annotation. The analysis revealed that almost 99% of the total number of genomic regions (which may cover multiple markers) are associated with one or two genes (Figure 40), with approximately 62% of genomic regions located between 0-500 kb downstream of a transcription start site (Figure 41).

The same marker list analysed for potential biological processes using the AmiGO Gene Onthology component of GREAT, confirmed that all the markers are involved in various stages of embryonic development, including: embryonic morphogenesis, sensory organ development, tissue development, pattern specification process, tissue morphogenesis, ear development, tube morphogenesis, epithelium development, chordate embryonic development and morphogenesis of an epithelium (each term has a clickable hyperlink to the relevant webpage in the digital version of this document).

Analysis of potential mouse phenotype associations confirmed that orthologous candidate markers were previously detected in mice models displaying abnormal morphology of the skeleton, head, viscerocranium and facial area, as well as specific malformations of the eye, ear, jaw, palate, limbs, digits and tail.

In terms of the molecular function, AmiGO revealed that candidate markers might be involved in a range of regulatory activities, including: protein dimerization activity, chromatin binding, regulatory region DNA binding, sequence-specific DNA binding RNA polymerase II transcription factor activity, sequence-specific distal enhancer binding activity, heparin binding, RNA polymerase II core promoter proximal region

sequence-specific DNA binding transcription factor activity involved in positive regulation of transcription, BMP receptor binding and transmembrane receptor protein serine/threonine kinase binding (each term has a clickable hyperlink to the relevant webpage). In summary, this study provides additional insight into the genetic context of targeted markers, confirming the craniofacial link of these SNPs.

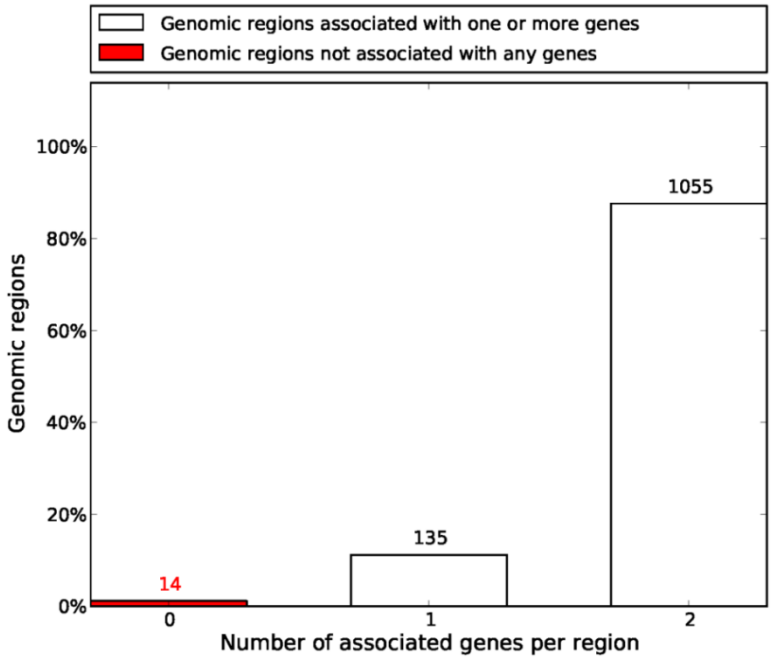


Figure 40. An illustration of candidate craniofacial markers, associated with genes. The majority of amplicons covered multiple markers and were associated with more than one gene.

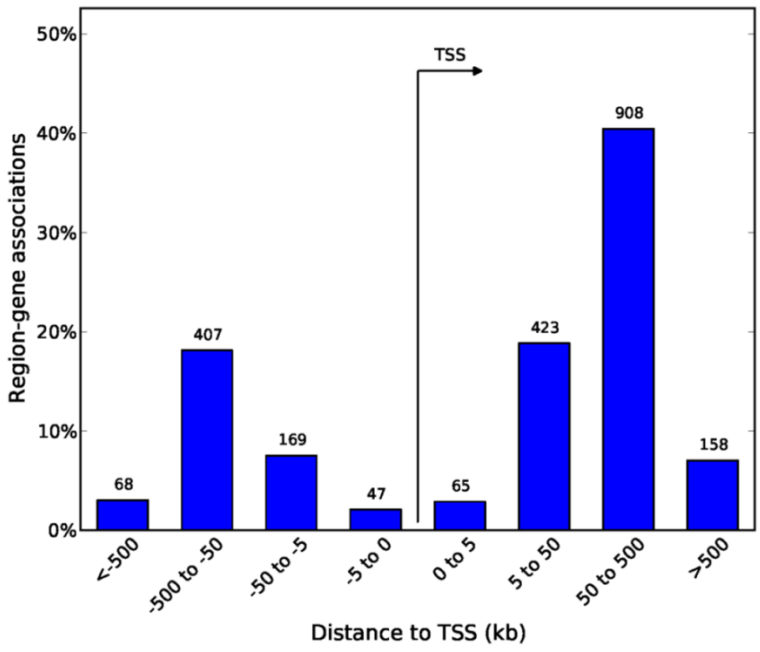


Figure 41. An illustration of genomic location of candidate markers in respect to the transcription start site.

3.5. Evaluation of the genotyping methods using GoldenGateTM assay on BeadExpress platform

As a part of a pilot study, the Illumina Golden Gate^{TM} technology was evaluated with a set of 96 pigmentation SNPs for prediction of skin, eyes and hair colour, as well as ancestry in pristine and suboptimal DNA samples on a BeadExpress platform. Specifically, the first study (Section 3.5.1) evaluated the performance of SNPs typing from a variety of intact and environmentally challenged DNA samples amplified with and without whole genome amplification (WGA) method [349]. The second study (Section 3.5.2), investigated the genomic association and subsequent prediction efficiency of a custom 96-SNP set. While the SNPs were originally chosen according to their association with pigmentation traits, they were also evaluated for potential prediction of ancestry in three US population groups. The results of these experiments were published or submitted for publication and provided as Sections 3.5.1 and 3.5.2 of this document. The results of both studies were included in the thesis of Dr Sheree Hughes-Stumm, submitted to the Bond University [350].

The association and prediction study provided an important opportunity for evaluation of both the genotyping and statistical analysis methods and established an initial foundation for subsequent study of externally visible traits and ancestry with more advanced genotyping technology (Chapter 5).

In addition, a subset of DNA samples and genotyped SNPs were overlapping between the GoldenGate assay and the Ion Torrent panel. This aspect made possible performing a validation study between two genotyping methods, as detailed in Section 3.6.3.

3.5.1. Initial evaluation of a 96-plex SNP panel for forensic analysis.

Avens Publishing Group
J Forensic Investigation
April 2013 Vol.:1, Issue:1

Initial Evaluation of A96-Plex Goldengate® Genotyping SNP Assay with Suboptimal and Whole Genome Amplified Samples

Open Access Research Article

Journal of Forensic Investigation

Sheree Hughes-Stamm[1*], Mark Barash[2], Kelly Grisedale[2] and Angela van Daal[2]

[1]*Department of Forensic Science, College of Criminal Justice, Sam Houston State University, Huntsville, TX, 77381, USA*
[2]*Faculty of Health Sciences and Medicine, Bond University, Gold Coast, QLD, 4223, AUSTRALIA*

Address for Correspondence
Sheree Hughes-Stamm, Department of Forensic Science, College of Criminal Justice, Sam Houston State University, Huntsville, TX, 77381, USA, Tel: (+1) 936 294 4359; E-mail: shereehs@shsu.edu

Submission: 04 September 2013
Accepted: 25 September 2013
Published: 27 September 2013

Keywords: Forensic; SNPs; Degraded DNA; WGA; MDA; GoldenGate

Abstract

A custom designed 96-plex GoldenGate® Genotyping single nucleotide polymorphism (SNP) assay was evaluated for performance with genomic samples (10-250ng template), whole genomic amplified (WGA) samples and environmentally challenged samples. The assay performed well with pristine genomic samples, reproducibly generating complete and accurate SNP profiles with lower (50ng) than the manufacturer's recommended amount of template (250ng).

Clinical and forensic samples often fall below the optimal concentration for direct SNP analysis. One proposed solution to this is to produce sufficient quantities of DNA prior to SNP typing by whole genome amplification. The results of this study show that WGA prior to SNP typing produced reliable results when the template was of high quality and quantity (≥10ng). However, SNP-typing of environmentally challenged skeletal samples produced poor results, both with and without WGA prior to SNP typing. The amplification bias inherent in the WGA process was significantly exaggerated with samples of low quality and quantity.

While these results suggest that SNP-typing using the Illumina GoldenGate® assay is not the solution for genotyping highly degraded samples, it canprovide an alternative means for some forensic DNA typing needs, such as paternity testing or typing reference samples for databasing.

Introduction

Due to the compromised nature of environmentally challenged and degraded samples, conventional short tandem repeat (STR) genotyping can result in a partial DNA profile or no profile at all. The forensic community has responded to this problem by designing mini-STR assays with smaller amplicons (<200bp) [1] and kits with more robust chemistry to overcome inhibition and increase sensitivity [2]. These innovations have resulted in a significant improvement in the genotyping success from challenging samples [3]. However, SNP analysis promises even better genotyping results from highly fragmented DNA samples [4]. The principle advantage of SNPs is their short amplicon length (potentially 45-55bp) [5].

Many different SNP technologies are available for genotyping. The majority of SNP genotyping assays can be classified based on the molecular mechanism employed: allele specific hybridization (ASO), primer extension, minisequencing, oligonucleotide ligation and invasive cleavage [6]. Detection methods for product analysis include fluorescence, luminescence and mass spectrometry. When deciding which platform to employ for clinical or forensic work, the three main considerations are; 1) the amount of DNA template required, 2), throughput capacity and 3) multiplexing capabilities. For this study, the Illumina Veracode GoldenGate Genotyping Assay was chosen for its multiplexing capacity and throughput (96 SNPs in a single well of a standard 96-well microplate).

Forensic samples often present as degraded and/or with limited amounts of DNA template. While these low amount or quality samples can often be successfully genotyped using STR and mini-STR analysis, SNP typing generally requires greater amounts of starting template. Samples below the minimum concentration required for robust SNP analysis, can potentially be amplified by whole genome amplification (WGA) to produce sufficient quantities of DNA. WGA of low amounts of DNA (50pg) and highly degraded samples has been shown to improve STR-typing [7] and SNP-typing [8]. WGA would also allow for additional testing and subsequent archiving. It is crucial to successful SNP genotyping that there is even amplification of the two alleles at each locus. It is also important that all regions of the genome containing SNP loci are amplified to approximately the same levels. One of the limitations of the WGA process is preferential amplification of some genomic regions and therefore production of an unfaithful representation of the genome. This most commonly results from the under-amplification of certain regions of the genome such as telomeres and centromeres [9,10] or regions containing repetitive elements [11] and regions of higher than average G-C content [12]. This need not be limiting because it could be overcome by judicious selection of SNP loci from regions known to be well amplified. It is more problematic to overcome imbalance that resultsfrom differential amplification at a particular locus or region resulting in allelic imbalance or dropout. One study found that use of a specific type of WGA termed multiple displacement amplification (MDA) for SNP typing via the GoldenGate® assay was feasible, but warranted caution regarding SNP selection based on SNP distance to telomere and local G-C content [13].

In addition to the standard MDA kit protocol, two modified methods of the same chemistry were investigated to determine whether either of these techniques would reduce the allelic imbalance

Citation: Hughes-Stamm S, Barash M, Grisedale K, van Daal A. Initial Evaluation of A96-Plex Goldengate® Genotyping SNP Assay with Suboptimal and Whole Genome Amplified Samples. J Forensic Investigation. 2013;1(1): 8.

effects of WGA, and therefore produce better quality SNP results than the standard protocol. The first method involved splitting and re-pooling the WGA reaction prior to SNP testing. This approach is similar to that used for low copy number DNA typing, except the reaction is split rather than the DNA sample. We hypothesised that splitting the total reaction volume into multiple aliquots during amplification and then re-pooling product prior to SNP typing may balance out any amplification bias generated within each individual MDA reaction. The second method subjected the sample to heating and cooling cycling conditions. As SNP targets are relatively short (~120bp), the long hyperbranching amplicons generated by the Φ29 DNA polymerase used during MDA may not be required, and may in fact contribute to the amplification bias. The hypothesis was that primers are forced off the template and back on to another random site in an attempt to generate multiple initiation events and create a more even coverage of the genome.

This study forms an initial evaluation of an Illumina GoldenGate 96-plex SNP assay on the Illumina BeadExpress platform for DNA samples, including whole genome amplified, degraded and environmentally challenged samples. The markers for this assay were chosen for their known association with normal human pigmentation (manuscript in preparation). The ability to predict external visible characteristics such as hair, eye and skin colour from a DNA sample could be of great assistance to law enforcement agencies as investigative leads.

Materials and Methods

Bone samples

Human bone samples (tibial plateau) (n=42) were exposed to one of five environmental insults: 1) saltwater submersion for 6 months, 2) freshwater submersion for 2 months, 3) burial for up to 24 months, 4) surface exposure for up to 24 months and 5) heat exposure of low (~350°C), moderate (~550°C) and high (~700°C) temperatures in a crematorium oven for up to 30 min (Figure 1). The surface of the bone was removed by sanding with a rotary power tool (Dremel Stylus™, USA) and broken into small pieces using a sterile chisel or cutting disc. Bone chips were treated with a wash series consisting of 4 steps; 20% bleach, 2 washes of sterile water and 100% ethanol each incubated for 5 min (900rpm at 22°C) then dried. Bone chips were ground into a fine powder using a liquid nitrogen freezer mill (6770 SPEX, USA). The freezer mill cycle conditions consisted of a ten minute pre-cool, 2x one minute crush, 2x one minute cool.

DNA extraction

DNA was extracted from whole blood (n=15) using the QIAamp® Blood Maxi Kit (Qiagen, Hilden, Germany), and from buccal swabs (n=42) using the QIAGEN EZ1 Tissue kit on the QIAGEN EZ1 robot as per manufacturer's instructions.

Bone powder (0.5-0.8g) was digested using a modified total demineralisation method and extracted using the QIAGEN Blood Maxi kit [14,15]. Powder was incubated in 10mL ATL Buffer (QIAGEN), 5mL 0.5M EDTA, 150µL Proteinase K (20mg/mL) and 200µL 1M DTT at 56°C for 24 h in a shaking incubator. An additional 5mL 0.5M EDTA was added and returned to the shaking incubator at 56°C for a further 24 h. Subsequently, 15mL AL buffer (QIAGEN) and 150µL Proteinase K (20mg/mL) was added and incubated at 70°C for 1 h in a shaking incubator. Samples were centrifuged at 1000g for 5 min and the supernatant was mixed with 15mL 100% Ethanol and added to QIAGEN Blood maxi spin columns. The columns were centrifuged for 3 min at 2000g, washed with 10mL AW1 buffer and centrifuged again. The filter was washed with 10mL AW2 buffer and centrifuged at 2000g for 3 mins. Residual AW2 buffer was removed by further centrifugation at 2000g for 10 minutes. DNA was eluted by adding 3mL AE buffer preheated to 72°C and centrifuged at 2000g for 3 mins. The elution process was repeated by running eluate back through the filter for maximal concentration. All samples were concentrated by centrifugation for 3 minutes at 4000g in Amicon Ultra-4 columns (Millipore) to a final volume of 200-400µL. Samples were further concentrated using a 30 min centrifugation in a Speedvac (Thermo scientific) at the medium heat setting (<100µL final volume).

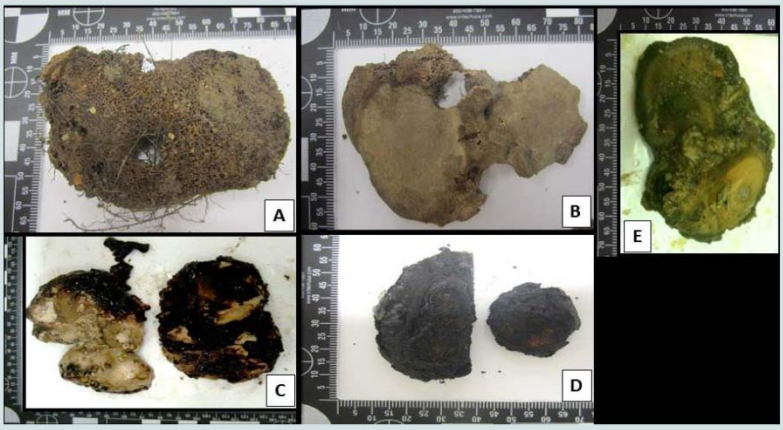


Figure 1: An example of bone samples, treated by each of the five environmental insults: (A) burial for 24 months,(B) surface exposure for 24 months,(C) freshwater submersion for 2 months, (D) heat exposure of~700°C for 20 min and (E) saltwater submersion for 6 months.

DNA Quantification

The DNA extracts were quantified using a real-time quantitative PCR (qPCR) assay. This assay amplified a 63bp region of the *hTERT* locus. The primer sequences were 5'-CAGCTTCCTTCGTTGAGGAG-3' (forward primer) and 5'-GAACAGCAATGACAGGCAGA-3' (reverse primer) at a final concentration of 200mM. The assay was performed on a Rotor-Gene 6000 (Qiagen) real-time thermocycler in a 25µL reaction volume using SensiMix HRM Master Mix (Bioline). The three-step qPCR protocol consisted of an initial 15 minute 95°C *Taq* DNA polymerase activation step, followed by 40 cycles of 15 seconds of denaturation (95°C), 10 seconds of annealing (60°C) and 10 seconds extension (72°C). HMW human genomic male DNA of known concentration (Promega, Madison, WI) was used as a qPCR quantification standard (0.0254 - 25.4ng/µL). Standard curves with good linearity (R^2 values above 0.99) were accepted for analysis. No template controls (NTCs) were included to monitor contamination during PCR.

Whole genome amplification (WGA)

DNA (n=15) was extracted from whole blood using the QIAamp® Blood Maxi Kit (Qiagen, Hilden, Germany), as per manufacturer's instructions. DNA (10ng) was added to IllustraGenomiPhi V2 DNA Amplification kits (GE Healthcare, Buckinghamshire, UK) and processed in one of 3 ways:

MDA was performed according to specifications provided by the supplier (GE Healthcare) [16].

2. Split and pool. Reaction volume of 20µL was split into 4 aliquots of 5 µL each. Double volume (40µL) reaction split into 8 separate aliquots of 5µL each. Individual aliquots were subjected to isothermic amplification as per the standard kit protocol and were subsequently pooled together a prior to SNP analysis.

3. Cycling method. Samples were prepared as per the standard WGA kit protocol, but were subjected to a 30-cycle PCR-amplification of 30°C for 10 seconds and 40°C for 5 seconds in place of the isothermic incubation.

Neat WGA product (1µL, in triplicate) was added to each SNP-typing reaction.

SNP genotyping and analysis

SNP genotyping was performed using 10, 20, 50, 125 and 250ng genomic DNA extracted from whole blood (n=10) as well as from bone samples. The amount of DNA retrieved from each bone sample was variable (1-250ng/µL). However, for comparative purposes, the same volume of neat DNA extract (1µL) from each sample was added to the SNP reaction directly, and also to the WGA reaction prior to SNP-typing.

SNP genotyping was performed using the 96-plex GoldenGate® Genotyping Assay on the BeadExpress™ (Illumina, San Diego, USA). These 96 SNPs were chosen based on known association with human pigmentation for eye, hair and skin colour (Supplemental table 1). However, in this study, the assay was used solely to determine the success rate of SNP typing from highly degraded samples. Oligonucleotide pool assay (OPA) multiplex primer sets were designed and manufactured by Illumina for 96 SNPs (Electronic Supplement Table 1). Of the 96 SNP regions to be amplified, 94 were 121bp, and the other two 103bp and 93bp respectively. Each SNP was assigned a score (0-1.1) by the manufacturer according to how well it is expected to performtheoretically based on primer design parameters, with higher values indicating a greater likelihood of a successful SNP genotype assignment (Electronic Supplement Table 1).

Data analyses were performed using GenomeStudio™ Genotyping software (Illumina, San Diego, USA). Several indices were used to determine the performance of each SNP, the confidence of each SNP call, and the assay performance as a whole. The call rate for each SNP indicates the proportion of samples assigned a genotype at each locus. The GenTrain Score (GTS) indicates how well separated and tight the clusters of the different genotypes are for each SNP. GenCall Score is a quality metric that indicates the reliability of each genotype call (0-1). The 50% GenCall Score (p50 GC) for a particular locus represents the 50th rank for all GenCall scores for that locus.

SNPs were considered to be performing well if the GenTrain score was >0.3, the GenCall 50 was >0.4 and call rate was >90%. SNPs were considered to perform poorly if the call rate was 50-90%, but the GenTrain score was >0.3, and the GenCall 50 was >0.4.SNPs were excluded if the samples did not separate into well-defined clusters (GenTrain Score <0.3), if more than 50% samples showed <0.15 relative fluorescence unit (RFU), or if the overall call rate for that SNP was less than 90%.

Individual samples were excluded from analysis at a particular SNP if the RFU value was below 0.15 or the GenCall value was less than 0.25. Samples were considered good quality if the call rate was 98-100%, moderate if 95-97% and acceptable if 90-94%. All samples with call rates >90% were included for genotyping. Samples were excluded from analysis of the entire assay if the call rate was below 90%.

A miscall was assigned if an allele dropped out, or a wrong allele call was made compared to the reference. In the case of samples with degraded references, a consensus approach was employed with a genotype being called when seen at least twice in the triplicate analyses.The consensus approach involved calling a genotype when it was concordant with the reference and at least one other sample replicate. Sequencing was not performed to confirm genotyping accuracy.

Results and discussion

Sensitivity Study

Genomic DNA samples (n=10 individuals) were typed to determine the lower limit of accurate genotyping with the 96-plex GoldenGate® Genotyping Assay. Samples were run in duplicate or triplicate (138 data points) spread across three plates, with DNA

Table 1: Comparison of three WGA protocols.

Call rate (%)	Percentage of Samples (%) via Method		
	WGA Std	**WGA S/P**	**WGA Cycling**
98-100	**83.3**	**100**	**90**
95-97	0	0	10
90-94	16.7	0	0
<90 Excluded	0	0	0
Miscalls (% of total genotypes called)	0.3	0	0
No. Samples	12	9	10

n= 63 (3 biological samples, 7 technical samples per treatment).

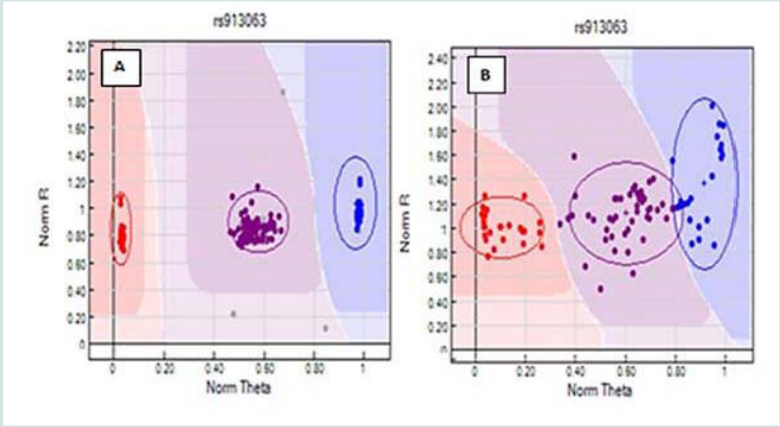


Figure 2: An example of the difference in clustering patterns between genomic and WGA samples

The same DNA samples were SNP-typed A) before and B) after MDA. The clustering of the three possible genotypes at SNP rs913063 is a typical example of the more diffuse clustering when samples were whole genome amplified prior to SNP-typing.
Graphic program used: Genome Studio (Illumina). Copied into Powerpoint

amounts of 10, 20, 50, 125 and 250ng. The reliability of SNP calls was consistently high across all amounts of DNA as demonstrated by p50GC values (0.68-0.74), which reflect the tightness of the genotype clustering. The assay showed reproducible and highly accurate results with as little as 50ng sample. The recommended input amount of DNA (250ng/sample) yielded the highest proportion of samples with 100% call rate. If samples are considered to perform well with a call rate threshold of >98%, then the assay was successful to a lower limit of 20ng. At 20ng template, >96% of all samples reached the SNP call rate threshold of 98%. However, with DNA amounts less than 50ng, genotyping miscalls were observed. At 20ng DNA the proportion of samples generating a 100% call rate dropped to 53.8% (96% of samples with a call >98% call rate) and low levels of miscalls were detected (0.041% of genotypes called). At 10ng, the call rate of samples within the 98-100% range dropped to 91.7% and a slightly elevated miscall rate of 0.05% of all genotype calls was observed. In addition, the slightly lower p50GC value (0.68 versus 0.74) shows that 10ng input of DNA produces a wider distribution of samples within each genotype cluster.

WGA

SNP typing of WGA samples routinely produced more diffuse clusters at more loci than gDNA samples (Figure 2). Compared to gDNA samples, WGA samples displayed almost twice the rate of miscalls (average 0.06% vs 0.11%), and a higher number of SNPs which performed poorly, or were excluded (8 vs 14 SNPs). This is likely due to the inherent amplification bias and/or stochastic effects from the WGA process. Fifteen SNPs were identified as requiring further evaluation (Electronic Supplement Table 1). Seven SNPs (rs3212355, rs3212359, rs1110400, rs2228479 rs3212361, rs1805008, rs6867641) performed poorly or required exclusion from both genomic and WGA sample analyses. The remaining eight SNPs showed poor performance with WGA samples but accurately genotyped with gDNA. One of the eight SNPs (rs6152) always performed poorly with WGA samples. This is most likely because this SNP is located in a GC rich region (~70%), relatively close to the centromere (of the X chromosome) [17].

The assay design process by the manufacturer involves each SNP being assigned a Final_Score (0 – 1.1) with Illumina recommending SNPs ranked <0.4 not be included in the multiplex assay [18]. However, eight separate runs of the SNP-panel demonstrated that this predictive score is not necessarily an accurate reflection of the actual performance and success of each SNP (Table 1). Nine of the 96 SNPs in the panel were assigned a Final_ Score of <0.4 (Electronic Supplement 1) by Illumina. Of these nine, only three were found to routinely fail and require exclusion from analysis, whilst five consistently performed well in the assay, and one has shown variable performance between plates. Conversely, although the majority of SNPs (90.6%) with Final_ Scores >0.4 performed well, it was not a guarantee of good performance. Within the group of seven SNPs which routinely performed poorly or required exclusion in this study, four had Final_ Scores>0.4.

Six of the seven SNPs requiring assay exclusion are located within the MC1R gene on Chromosome 16. This is likely due to the large number of SNPs located close to each other leading to problems with primer design. In addition, MC1R is telomeric in location, and has a higher (63% versus 40%) GC content than a typical human gene [19]. This difficulty was predicted with three of the six MCR1 SNPs allocated Final_Scores below the recommended threshold (0.4), and the relatively low scores of the other three SNPs (Final_Scores 0.4-0.6).

The SNP assay exhibited substantial run-to-run variability with respect to relative cluster positions and average RFU intensities (Figure 3). Although this made the use of genotype cluster positions at the same SNP across different runs more difficult and time consuming, accurate genotyping was still possible. The variance is thought to be primarily due to differences in sample quality, but differences in assay chemistry, batches and OPAs over time may also contribute. A

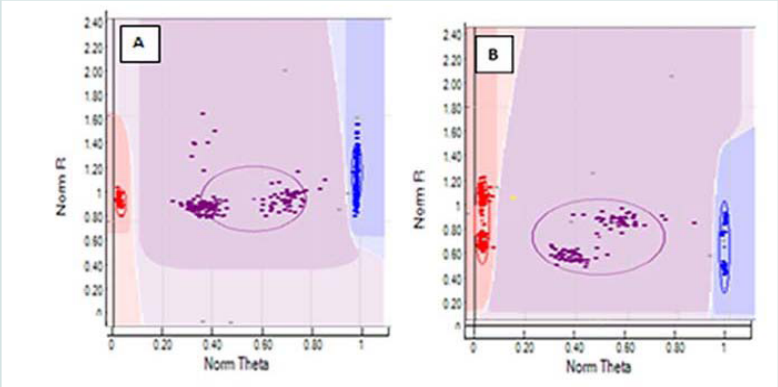


Figure 3: The observable run-to-run variability in clustering and RFU values

Two different SNP plates (n=90 each) combined for analysis. With both SNPs A) rs1042602, and B) rs1800404 two distinct populations can be seen. The heterozygote cluster in A) and all three genotype clusters in B) are evidence of inter-run variability. Note the theta shift between the heterozygote samples in image A, and the difference in RFU values between the two sets of samples in image B.
Graphic program used: Genome Studio (Illumina). Copied into Powerpoint

slight difference in general assay performance was observed between fresh and older (6-8 months) reagents in terms of RFU intensities and cluster patterns. However this was not thought to affect genotyping results. As expected, these observations were exaggerated in WGA sample plates due to the effects of amplification bias inherent in the WGA process.

Two modifications to the recommended protocol for the GenomiPhi V2 DNA Amplification kit were investigated using a small subset of samples (n=3 with 7 technical replicates) to seek an optimal method for WGA prior to SNP typing. Overall the standard WGA protocol resulted in the poorest SNP typing results compared to split/pool and cycling methods with the lowest call rate (83.3%) and the greatest percentage of miscalls (0.3%) (Table 1). Both alternate WGA methods resulted in 100% of samples having 95-100% call rates. However, it should be noted that the split and pool samples produced the highest proportion of samples with a perfect 100% call rate.

The slightly improved SNP-typing performance of WGA samples which have been split and pooled over those with the standard and cycling methods is most likely due to a more balanced amplification of the genome. These data support previous work showing that pooling of MDA product from two or three separate amplifications significantly reduced any allele amplification bias during WGA [20]. The number of aliquots per sample was investigated to assess whether amplification bias may be reduced with an increased number of sub-sample reactions.

WGA performed with pristine DNA (10ng) in eight sub-sample reactions compared to five aliquots generated significantly better SNP results (Table 2). This is measured by an increased percentage of samples (83.3% vs 64.6%) with call rates >98%, a decreased number of samples requiring exclusion (3.3% vs 8.3%) and less miscalls (0.08% vs 0.1%).

Bone Samples

SNP typing of highly degraded bone samples using this GoldenGate® assay gave poor results. Almost half of the samples (47.6%) required exclusion from the assay (Table 3). This poor performance was made significantly worse by WGA prior to SNP analysis with 71.4% of samples excluded. Although the split and pool

Table 2: Comparison of two Split and Pool methods (10ng input).

Call rate (%)	WGA Split and Pool Method (% of samples)		Genomic DNA (% of samples)
	5 aliquots	8 aliquots	
98-100	**64.6**	**83.3**	**81.3**
95-97	18.8	10	18.8
90-94	8.3	3.3	0
<90 Excluded	8.3	3.3	0
Miscalls (% of total genotypes called)	0.1	0.08	0
No. Samples	48	30	16

n= 90 (15 biological samplesper treatment, in duplicate)

Table 3: SNP-typing results from Bone samples.

Call rate (%)	Genomic (% of samples)	WGA treated (% of samples)
98-100	**30.9**	**11.9**
95-97	16.67	**11.9**
90-94	4.8	4.8
<90 Excluded	47.6	71.4
Average Call Rate	**78**	**55.7**
Median Call Rate	**93.1**	**77.6**
Miscalls (% of total number genotypes)	1.6	1.5
Unable to assign a genotype (% of total number genotypes)	1.2	0.7
Av p50 GC	**0.56**	**0.44**
No. Samples	42	42

WGA method was shown to be more successful with intact DNA samples in this study, an initial trial of these methods with highly degraded samples demonstrated that the non-pooled manufacturers recommended protocol yielded better results (data not shown), and was therefore used with all of the bone samples. A substantial difference in the average GeneCall scores was observed between the genomic bone and WGA bone samples (p50 GC: 0.56, 0.44 respectively) (Table 3). This indicates that the reliability of allele calls when they are made is generally low for the degraded bone samples and even less when those degraded samples are whole genome amplified prior to SNP typing. Although the MDA-based GenomiPhi method used in this study is one of the most commonly used WGA kits for forensic genetics, a more recent study has shown that other WGA methods such as GenomePlex perform better with degraded DNA (200 bp), and yield more complete SNP profiles [8].

Variation in overall performance was observed between the types of environmental exposure (Figure 4). Using the call rate as the primary measure of success, the samples exposed to heat yielded the most complete SNP profiles with all three heat treatments producing average call rates above 98% (Figure 4a). This may be explained by the fact that sufficient amounts of good quality DNA were able to be extracted from these samples. As a result, it was possible to input the optimal 50-250ng template to the SNP assay. These bone samples were extremely charred on the outside, but the osseous tissue internally was well protected and therefore DNA quality was preserved within that tissue.

Freshwater samples decomposed much faster than saltwater samples showing extensive colonisation by microbes and algae. As a result these bones were incubated for only two months compared to six months for the saltwater samples. Even with a longer exposure period, the saltwater samples yielded higher SNP call rates than the freshwater samples (99.4% vs 82.2% respectively) (Figure 4b). It is known that saltwater slows the rate of cadaveric decomposition by suppressing bacterial growth [21]. These SNP data suggest that the saltwater environment may also preserve DNA within bone tissue.

The average genotyping success of surface samples was consistent for up to one year exposure (approximately 95%). However a 15% decrease in call rate was seen in samples with 18 and 24 months exposure (Figure 4c), which brings the rate below 90% and therefore not considered reliable for genotyping. The buried samples generated the poorest and the most variable SNP profiles (0-90% call rates) (Figure 4d).

Regardless of the mode of environmental insult, genotyping was universally more successful without WGA prior to SNP-typing. These data suggest that WGA (or the methods investigated in this work)

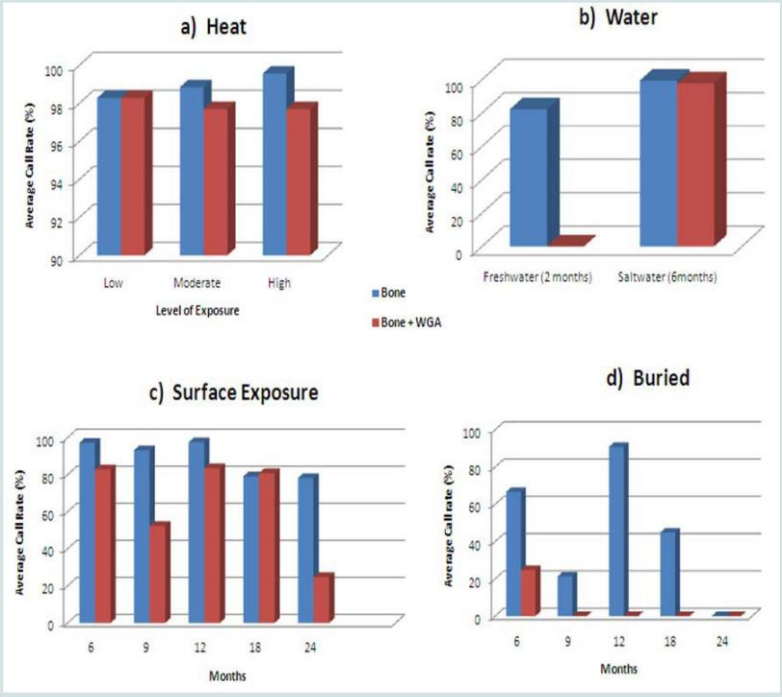


Figure 4: SNP typing success of environmentally challenged samples with and without WGA

The average SNP call rate of a) heat treated, b) freshwater and salt water, c) surface and d) buried bone samples. Each bone sample was directly SNP-typed, or whole genome amplified prior to SNP-typing.
Graphic program used: Graphs created in excel.Copied into PowerPoint.

of low DNA quality and quantity creates template DNA which is unsuitable for reliable SNP typing using the Veracode GoldenGate® Genotyping SNP Assay on the BeadExpress platform. However, it must be noted that the SNP panel in this study used custom amplicons 93bp and 103bp in length, a result of the manufacturer's design process. Better results with degraded samples may be expected if smaller targets were included in the panel and would therefore be an important consideration in multiplex design for any future SNP panel for forensic use.

Conclusion

The custom designed 96-plex Veracode GoldenGate® Genotyping SNP Assay on the BeadExpress (Illumina) was evaluated for performance with optimal genomic samples (250ng template), samples of low quantity (10, 20, 50 and 125ng), whole genomic amplified samples, and degraded and environmentally challenged samples. Overall, the assay performed well with pristine genomic samples and sensitivity studies showed that the assay is quite tolerant to less DNA template than recommended by the manufacturer. The SNP multiplex reproducibly generated complete and accurate SNP profiles for genomic samples of 50, 125 and 250ng.

Forensic and clinical samples may fall below the optimal concentration for direct SNP analysis. The low amounts of DNA for analysis may also prevent multiple testing and archiving for future use. One proposed solution to this problem is whole genome amplification as a means to produce sufficient quantities of DNA prior to SNP typing. However, the results of this study show that although the application of WGA prior to SNP typing produced reliable results if the template for WGA is of high quality and quantity (10ng), the clustering of samples at each locus is more diffuse, requiring more scrutiny during data analysis.

As DNA database and forensic samples are often limited and in low concentrations, it is vital to input minimal DNA into each reaction. However, as is the case for many degraded samples, this option is not always possible. While forensic STR analysis can be conducted on sub-nanogram amounts of DNA, SNP technology is yet to reach that level of sensitivity. Inclusion of a WGA step is therefore one potential means to allow SNP analysis of nanogram or sub-nanogram amounts of evidentiary DNA samples. However, due to the superior performance of the SNP typing of non-WGA samples, it is considered preferable to genotype samples without WGA when possible.

WGA of degraded samples prior to SNP-typing produced poor results. The amplification bias inherent in the WGA process is significantly exaggerated with samples of low quality and quantity. This research suggests that neither direct SNP-typing using the Illumina GoldenGate® assay nor MDA-based WGA prior to genotyping is the solution for genotyping highly degraded samples. Relatively large amounts (>50ng) of good quality DNA are required for accurate SNP analysis. However alternate WGA methods, SNP arrays and next generation DNA sequencing platforms may generate improved results. These techniques may also provide the opportunity to multiplex much larger sets of markers from smaller amounts of DNA (<20ng).

A substantial reduction in SNP typing performance was observed with samples of low quality and quantity. The 96-plex SNP panel genotyped using the GoldenGate® technology was not robust with sub-optimal samples such as low DNA amounts; whole genome amplified or degraded samples. The combination of time-consuming bench work, low call rates, unacceptably high miscall rates and complex (and often subjective) data analysis makes this platform unsuitable for the analysis of poor quality samples. However, in situations such as reference databasing, paternity testing and clinical diagnoses when adequate amounts of high quality DNA are available, the GoldenGate® technology could provide reliable SNP-typing platform.

References

1. Wiegand P and Kleiber M (2001) Less is more--length reduction of STR amplicons using redesigned primers. Int J Legal Med 114: 285-287.
2. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, et al. (2006) Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. J Forensic Sci 51: 64-75.
3. Welch L, Gill P, Tucker VC, Schneider PM, Parson W, et al. (2011) A comparison of mini-STRs versus standard STRs--results of a collaborative European (EDNAP) exercise. Forensic Sci Int Genet 5(3): p. 257-258.
4. Fondevila M, Phillips C, Naverán N, Cerezo M, Rodríguez A, et al. (2008) Challenging DNA: Assessment of a range of genotyping approaches for highly degraded forensic samples. Forensic Science International: Genetics Supplement Series 1: 26-28.
5. Kidd KK, Pakstis AJ, Speed WC, Grigorenko EL, Kajuna SL, et al. (2006) Developing a SNP panel for forensic identification of individuals. Forensic Sci Int 164: 20-32.
6. Sobrino B, Brion M, Carracedo A (2005) SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int 154: 181-194.
7. Ballantyne KN, van Oorschot RAH, Mitchell RJ (2007) Comparison of two whole genome amplification methods for STR genotyping of LCN and degraded DNA samples. Forensic Sci Int 166: 35-41.
8. Maciejewska A, Jakubowska J, Pawlowski R (2013) Whole genome amplification of degraded and nondegraded DNA for forensic purposes. Int J Legal Med 127: 309-319.
9. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U.S.A 99: 5261-5266.
10. Panelli S, Damiani G, Espen L, Micheli G, Sgaramella V (2006) Towards the analysis of the genomes of single cells: further characterisation of the multiple displacement amplification. Gene 372: 1-7.
11. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, et al. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. Nucleic Acids Res 32: e71.
12. Han T, Chang CW, Kwekel JC, Chen Y, Ge Y, et al. (2012) Characterization of whole genome amplified (WGA) DNA for use in genotyping assay development. BMC genomics 13: 217.
13. Cunningham JM, Sellers TA, Schildkraut JM, Fredericksen ZS, Vierkant RA, et al. (2008) Performance of amplified DNA in an Illumina GoldenGate BeadArray assay. Cancer Epidemiol Biomarkers Prev 17: 1781-1789.
14. Davoren J, Vanek D, Konjhodzic R, Crews J, Huffine E, et al. (2007) Highly effective DNA extraction method for nuclear short tandem repeat testing of skeletal remains from mass graves. Croatian medical journal 48: 478-485.
15. Seo SB, Zhang A, Kim HY, Yi JA, Lee HY, et al. (2009) Technical note: Efficiency of total demineralization and ion-exchange column for DNA extraction from bone. Am J Phys Anthropol 141:158-162.
16. (2006) Healthcare G, illustra GenomiPhi V2 DNA Amplification Kit. , in Product Web Protocol.
17. (NCBI) NCfBI. dbSNP Database.
18. (2009) Illumina I, Technical note: Designing custom GoldenGate genotyping assays.
19. White DL and Rebbeck TR (1999) Using MasterAmp™ PCR PreMixes to Optimize Analysis of High GC Content Genes: PCR Amplification of the Melanocortin-1 Receptor Gene. Epicentre Forum.

Citation: Hughes-Stamm S, Barash M, Grisedale K, van Daal A. Initial Evaluation of A96-Plex Goldengate® Genotyping SNP Assay with Suboptimal and Whole Genome Amplified Samples. J Forensic Investigation. 2013;1(1): 8.

20. Rook MS, Delach SM, Deyneko G, Worlock A, Wolfe JL (2004) Whole genome amplification of DNA from laser capture-microdissected tissue for high-throughput single nucleotide polymorphism and short tandem repeat genotyping. Am J Pathol 164: 23-33.

21. Rodriguez III WC (1997) Decomposition of buried and submerged bodies, in Forensic Taphonomy, William D Haglund , Marcella H . Sorg, Editors , CRC Press: Boca Raton.

Acknowledgements

The authors acknowledge the generous support provided by the National Institute of Justice (NIJ), the Technical Support Working Group (TSWG) of the United States Government National Interagency Research and Development Program for Combating Terrorism. In addition, substantial in-kind support was provided by Illumina, Inc.

All appropriate ethical approvals have been granted for this project (Bond University Human Research Ethics Committee approval RO-510).

3.5.2. Evaluation of a 96-plex phenotypic SNP panel for the prediction of ancestry, eye, skin and hair colour in the three major US population groups.

Order of Authors:	**Sheree Robyn Hughes-Stamm, Ph.D.**
	Mark Barash, MSc., BSc.
	Philipp E Bayer, BSc. (Hons)
	Grisedale Kelly, MSc. BSc.
	Wenjie Lui, Ph.D., BSc. (Hons)
	Angela van Daal, Ph.D., BSc. (Hons)
Abstract:	This study evaluated the power of an interacting subset of 96 SNPs, to predict the ancestry and various pigmentation traits of the source of a DNA sample using a custom GoldenGate® assay. Statistical models were built for the prediction of hair, eye and skin colour. In addition, statistical models were built to test the prediction of the three main US population groups (African American, Caucasian and Hispanic). A blind test of each model using 100 samples showed the following accuracies for each of the traits investigated: 92% - 96% for ancestry, 73% - 94% for hair colours, 78% - 97% for skin colours and 83% - 87% for eye colours. This study forms an initial evaluation of this custom SNP assay for forensic identification purposes.
Author Comments:	Evaluation of a 96-plex phenotypic SNP panel for the prediction of ancestry, eye, skin and hair colour in the three major US population groups. Sheree Hughes-Stamm1,2,*, Mark Barash1,*, Philipp E. Bayer3,4, Kelly Grisedale1, Wenjie Lui1, Angela van Daal1 * **Hughes-Stamm and Barash are joint first authors** 1 Faculty of Health Sciences and Medicine, Bond University, Gold Coast, QLD, 4223, AUSTRALIA 2 Department of Forensic Science, College of Criminal Justice, Sam Houston State University, Huntsville, TX, 77381, USA 3 Applied Bioinformatics Group, School of Agriculture and Food Sciences, Faculty of Science, University of Queensland, Brisbane, AUSTRALIA 4 Australian Center for Plant Functional Genomics, School of Agriculture and Food Sciences, University of Queensland, Brisbane, AUSTRALIA

Abstract

This study evaluated the power of an interacting subset of 96 SNPs, to predict the ancestry and various pigmentation traits of the source of a DNA sample using a custom GoldenGate® assay. Statistical models were built for the prediction of hair, eye and skin colour. In addition, statistical models were built to test the prediction of the three main US population groups (African American, Caucasian and Hispanic). A blind test of each model using 100 samples showed the following accuracies for each of the traits investigated: 92% - 96% for ancestry, 73% - 94% for hair colours, 78% - 97% for skin colours and 83% - 87% for eye colours. This study forms an initial evaluation of this custom SNP assay for forensic identification purposes.

Keywords
Forensic DNA analysis, Forensic SNPs, Human Appearance, Forensic Phenotyping, ancestry, Golden Gate assay

Introduction

The most obvious descriptors of an individual's appearance are colouring, height and facial features [1]. All are highly heritable [2, 3] and therefore the genetic basis of such traits may be exploited for possible forensic intelligence. Hair, eye and skin colour are visual traits which are fundamental eyewitness descriptors. Many studies have focused on the genetic factors influencing pigmentation of skin [4-7], hair [8-10] and eye colour [11-18]. In addition to these external visual characteristics (EVCs), the assignment of ancestry to a person (or DNA sample) can provide important information. The ability to determine EVCs and ancestry from DNA can assist in solving missing person cases, identifying mass disaster victims, providing intelligence information for crime scene samples when no suspect is identified, or giving a 'face' to historical skeletal remains.

Many genes and DNA polymorphisms are known to be associated with variations in human pigmentation. Numerous SNPs that are strongly associated with external visible characteristics have been identified [19]. Key genes found to be predictive of hair, skin and/or eye colour include the HERC2, MC1R, OCA2, SLC24A5, SLC45A2 (MATP), ASIP, IRF4, MYO5A and TYR genes [20-22, 17]. As many pigmentation SNPs vary across population groups and geographical areas, it may also be possible to determine the ancestry of an individual based on selected pigmentation SNPs. Recent studies have focused on combining both sets of markers into a single panel of predictive SNPs [23-25]. This approach is known as forensic DNA phenotyping [26].

This pilot study investigated whether the custom GoldenGate® 96-plex pigmentation SNP panel was able to predict eye, skin and hair colour as well as ancestry (three major US populations) from a DNA sample. As pigmentation phenotypes vary across population groups and geographical areas, it might be expected that pigmentation SNPs may also be predictive of ancestry to some degree.

Materials and methods

Samples

DNA samples (n=917) were extracted from whole blood or buccal swabs using one of three methods; EZ1 Blood or Tissue kit (Qiagen, Hilden, Germany), DNA IQ (Promega, Madison, US) or PrepFiler (Life Technologies). The amount of DNA was determined by quantitative Real Time PCR (qPCR) using the Quantifiler Human Quantitation kit (Life Technologies). Quantitative Real Time PCR was performed on either Rotor-Gene 6000 (Qiagen), Bio-Rad CFX 96 (Bio-Rad, Gladesville, Australia) or 7500 Real-Time (Life Technology) thermal cyclers.

DNA samples (n=552) had eye, hair, skin colour and ancestry data recorded. An additional 365 DNA samples had only ancestry data recorded and were derived from the three major US population groups: Caucasian, African American and Hispanic. However, all African American samples were further assigned pigmentation data for dark skin, black hair and brown eyes.

SNP-typing

SNP genotyping was performed using the 96-plex GoldenGate® Genotyping Assay and the BeadExpress™ instrument (Illumina, San Diego, USA) as per manufacturer recommendations. Input amounts of DNA ranged from 10-250ng/sample (in duplicate or triplicate). The majority of samples (n=857) were typed directly using 100-250ng of DNA. A subset (n=60) of DNA samples were whole genome amplified (WGA) using the Illustra GenomiPhi V2 DNA Amplification kit (GE Healthcare, Buckinghamshire, UK) prior to genotyping. The manufacturer's recommended protocol for isothermic amplification was modified such that DNA (10-80ng) was added to a double reaction volume (40µL) and then split into 8 aliquots of 5µL for amplification. Individual aliquots were subjected to isothermic amplification as per the standard kit protocol. Aliquots were then pooled prior to SNP analysis. Neat WGA product (1µL) was added per SNP reaction. Genomic DNA controls were included to check concordance of WGA treated samples. Two no template controls were included in each SNP-typing 96-well plate to monitor contamination.

SNP selection

Approximately 115 SNPs known to be associated with pigmentation phenotype were initially selected and submitted to Illumina for multiplex primer design using the Illumina® Assay Design Tool (ADT). Oligonucleotide pool assay (OPA) multiplex primer sets were designed and manufactured by Illumina for 96 of the SNPs (Table S2). Each SNP was assigned a score (0-1.1) according to how well it is expected to perform, with higher values indicating a greater likelihood of a successful SNP genotype assignment [27]. Illumina recommend avoiding SNPs with a Final Score <0.4, as inclusion may decrease the overall assay performance. Eight SNPs with scores below 0.4 were considered essential for inclusion in the forensic pigmentation SNP panel because of known strong associations with hair phenotypes.

Data analysis

Data analyses were performed using GenomeStudio™ Genotyping software (Illumina, San Diego, USA). Several indices were used to determine the performance of each SNP, the confidence of each SNP call, and the assay

performance as a whole. The call rate for each SNP indicates the proportion of samples assigned a genotype at each locus. The GenTrain Score (GTS) indicates how well separated and tight the clusters of the different genotypes are for each SNP. GenCall Score is a quality metric that indicates the reliability of each genotype call (0-1). The 50% GenCall Score (p50 GC) for a particular locus represents the 50^{th} rank for all GenCall scores for that locus respectively.
SNPs were considered to be performing well if the GenTrain score was >0.3, the GenCall 50 was >0.4 and call rate was >90%. SNPs were considered to perform poorly if the call rate was 50-90%, but the GenTrain score was >0.3, and the GenCall 50 was >0.4. SNPs were excluded if the samples did not separate into well-defined clusters (GenTrain Score <0.3), if more than 50% samples showed <0.15 relative fluorescence unit (RFU), or if the overall call rate for that SNP was less than 90%.

Individual samples were excluded from analysis at a particular SNP if the RFU value was below 0.15 or the GenCall value was less than 0.25. Samples were considered of a good quality if the call rate was 98-100%, moderate if 95-97% and acceptable if 90-94%. Samples were excluded from analysis of the entire assay if the call rate was below 90%. A correct genotype was determined by sample replicate concordance.

Statistical Analysis

Two types of statistical analyses were performed. The first identified the most strongly associated SNPs for each phenotypic trait (p-value <0.05), based on a compressed linear mixed model and using the open-source software package (GAPIT) developed for genome wide association studies (GWAS) [28]. The second statistical analysis was performed to reach the main goal of this study, which was to maximize the predictive power of a set of SNPs for each pigmentation trait and ethnicity. For this purpose, a forward stepwise conditional logistic regression was performed on a well-genotyped subset of samples (n=817) using PASW Statistics 18 (SPSS, Inc., 2009, Chicago, IL). In each step of the regression, all successfully typed SNPs were tested to see if any additional SNP improved the predictive power of the model by measuring the R^2 value of the predictive model. In this way, several SNPs were grouped inside a model to form a vector of alleles and handled as a predictor of each trait. Each step of the regression iterated over all alleles and included them in the model when the p-value for prediction of the phenotype was improved. However, it must be noted that R^2 values do not strictly apply to logistic regression analysis, and therefore a pseudo-R^2 (Nagelkerke) value was used to provide a general indication on the goodness-of-fit of each predictive model. A better indication is one based on a comparison of predicted and observed values from the fitted model [29]. In addition, low pseudo-R^2 values with logistic regression are the norm [29].

SNPs which had more than 5% of data missing were manually removed from PASW statistical analysis as logistic regressions cannot handle missing data well. The algorithm used for the logistic regression analysis removed samples if even one datum point (genotype or phenotype) used to generate the predictive model was missing.
As a binary logistic regression can only handle binary outcomes, all phenotypes were regrouped and renamed using a simple Python-script to reflect binary phenotypes. For example, the phenotype “eye colour” was split into all possible variations, and each variation was translated into binary (e.g. “blue eye colour” (1) and “not blue eye colour” (0); “brown eye colour” (1) and “not brown eye colour” (0) etc). For each of these binary phenotypes PASW was used to apply a logistic regression, and a predicted group as well as the probability of the predicted group were saved.
To verify PASW's results a compressed mixed linear model [30] as implemented in GAPIT with P3D/EMMAX [31] was used. A mixed linear model differs from a logistic regression as it incorporates normally distributed random effects

to alleviate the effects of noise in the data. Compression first clusters individuals into groups, which greatly speeds up the analysis as the sample size is decreased by the clustering.

One of the advances of the mixed linear model against the binary logistic regression as it fares better when it comes to confounding due to population structure [32]. A mixed linear model can also handle missing data better compared to logistic regression. However, the interactions between SNPs are not measured, only the predictive power of single SNPs is computed. To ascertain the validity of the binary logistic regression models' predictions for each possible phenotype, 100 samples were randomly removed by a small Python script. On each of these datasets another binary logistic regression was run, as a logistic regression can predict the phenotype even without an actual phenotype-measurement being present. Another Python script was used to count the number of false positive and false negative predictions.

Results and discussion

This study was undertaken to assess the potential of the 96-plex SNP panel to predict hair, eye and skin colour, and ancestry. Individuals (n=915) from various population groups (primarily Caucasian, African American and Hispanic) were genotyped. These samples varied in DNA quantity and quality. Forward stepwise conditional logistic regression was used to build predictive models for each phenotype and ancestry. The logarithm used to generate the predictive models was not tolerant of missing data. As a result, samples were removed even if only one datum point (genotype or phenotype) used to generate the predictive model was missing. Therefore, if a model requires nine SNPs, samples must have a genotype for all nine SNPs, and the phenotype data for that trait assigned. This led to ~20% of samples in each trait failing to meet the strict requirements for inclusion in the statistical modelling (Table 1). In the following section, the results of the binary logistic regression are discussed first, followed by a discussion of the results gathered by the mixed linear model.

Due to the variation in DNA quality and quantity of the samples used, and the notable run-to-run variability of the assay, the completeness of SNP profiles obtained was inconsistent. This variability resulted in the number of SNPs being excluded in each run ranging from 1 to 12 of the 96 SNPs during the data quality control and clean-up process prior to the statistical analysis. The major influences on the performance of the GoldenGate® Genotyping assay were sample quality and quantity (manuscript in preparation).

Because logistic regression is not tolerant to missing data, SNPs with more than 5% of data missing (more than 5% of samples were not assigned a genotype for that particular SNP) were disqualified from the logistic regression. This threshold resulted in 24 of the 96 SNPs (25%) being removed. Three SNPs (rs2296151, rs41273937 and rs11547464) were also removed from analysis because they were monomorphic.

Eye Colour prediction

The 96-plex SNP panel was evaluated to identify SNPs which may be predictive of blue, brown, green and hazel eye colours. Statistical models were generated for the prediction of all four eye colours (Fig 1). The predicted accuracies ranged from 83-86% using three SNPs for predicting green eyes and six SNPs each for the other three colours (Table 2).

Interestingly, even though the predictive accuracy of all colours was similar, the statistical models for blue and brown eye colours were much more robust than the other two colours (R^2 values 0.6 versus 0.2) (Fig. 1). The predicted accuracy of blue and brown eye colour was consistent with other studies [33, 34, 14, 17]. The relatively low R^2 values in this study may be explained by the small sample size used. Additionally, the mix of ancestries in the group of hair colours led to confounding effect as individuals with different ancestral backgrounds share hair colours, leading to different SNPs being associated at the same time with hair colour.

Of the 31 SNPs reported in key studies to be most strongly associated with eye colour variation, only 19 were included in the 96-plex panel tested in this study, and a further 5 were excluded from statistical analysis (Electronic Supplement Table S3). The fact that less than half of the SNPs currently known to be most predictive for eye colour were included in this study may explain the relatively lower predictive accuracy and R^2 values obtained than seen in those previous works [34, 14, 17]. Five of the six SNPs reported to be key predictive indicators of eye colour (included in the IrisPlex

[35] and HIrisPlex [18] assays) were included in the 96-plex assay, but one (rs1800407) was excluded due to poor performance.

The rs12913832 and rs1129038 SNPs in the HERC2 gene have been reported as the most informative SNPs for eye colour [36, 12, 37], with rs12913832 being ranked as the single highest predictor [38, 17]. In fact rs12913832 has been reported to have strong predictive power for human pigmentation in general [8, 20, 22, 39].

Both SNPs (rs1129038 and rs12913832) are included in the same "h-1" haplotype spanning 166kB on chromosome 15, found in homozygous state (C allele, [40]) in 97% of Caucasian individuals with blue eyes [41]. Although these SNPs were the two most strongly associated with blue and brown eye colours ($p < 0.0001$) in this study (Electronic Supplement Table S4), the algorithm used for predictive models selects only one of the two SNPs. These two SNPs are linked, and therefore a reasonable approach would be to use one of these as a tag SNP for predictive purposes. One of these SNPs, rs12913832 was included in the predictive models for hazel and green eyes, whilst the other, rs1129038 was used to predict blue and brown eyes.

Prediction using a broader colour category such as 'blue eyes' versus 'not blue eyes', and 'brown' versus 'not brown' may also improve the accuracy rates for eye colour. In this study, that approach decreased the predicted accuracy for blue and brown eyes (blue; 83% to 79% and brown; 86% to 74%), but increased the accuracy predicted for 'not blue' and 'not brown' eyes (blue; 83% to 85% and brown; 86% to 93%).

Hair Colour

The ability of the 96-plex SNP panel to predict four common shades of hair colour (blonde, brown, red and black) was investigated. The prediction of hair colour was associated with the same genes contributing to eye and skin colour (OCA2, HERC2, SLC24A5, SLC45A2, MC1R and TYR) [42, 18], while the SNPs within the MC1R gene were found to be specifically associated with red hair and fair skin [43, 36, 44]. The predictability of hair colour has been reported with mixed levels of accuracy, being generally reported as either lower (<80%) than eye colour predictions [36], or equal to those for eye colour (>90%) [8].

A set of 10 SNPs (Electronic Supplemental Table S6) was found to be a highly predictive (94%) and robust model (R^2 value = 0.835) for the prediction of black hair (Fig. 2). A comparably high accuracy was predicted for red hair using five SNPs (Supplemental Table 6) despite no SNPs being identified as strongly associated ($p<0.05$) with red hair in this study (Electronic Supplemental Table S4). This model also showed a weaker association with a R^2 value of 0.311. These low values and lack of SNPs showing strong association is not surprising, as no MC1R SNPs were included in the analyses due to poor performance. Sensitivities for predicting blonde and brown hair had lower rates of accuracy with 88% and 73% respectively (Fig 2). Interestingly, no significant SNPs ($p<0.05$) for either blonde or brown hair association were identified in this study (Electronic Supplemental Table S4).

Several studies [8, 45, 43, 36, 44, 18] have identified SNPs significantly associated with hair colour (Electronic Supplement Table S3). Of those 46 SNPs, 17 were not included in the 96-plex assay, 6 were removed from analysis and another 5 were found to be not predictive of the pigmentation phenotypes (Electronic Supplement Table S6). Of the remaining 18 SNPs, 6 were found to be associated ($p<0.05$) with hair colour (Electronic Supplement Table S6). Three of those six SNPs (rs11547464, rs1805008 and rs1800407) were excluded from predictive analyses due to >5% missing

data, and the remaining three SNPs were included in the predictive models for skin colour, eye colour and/or ancestry. The 10 SNPs which were used to predict hair colour were, in general, not closely associated with the same hair colour as reported in the literature [8, 36]. As only a few of the known SNPs strongly associated with hair colour were used to generate the predictive models, the actual accuracy of the prediction models may be expected to be relatively low.

Skin Colour

Statistical models were generated to predict four shades of skin colour (fair, average, olive and dark) in this study (Fig. 3). The model for dark skin used 20 SNPs and predicted with 97.1% accuracy. The SNP model for dark skin was also the strongest with an R^2 value of 0.917. The next most informative model was the average skin with a predicted accuracy of 89.4%. The fair and olive skin predictions were quite low, with 78% and 79% accuracy respectively. However, the pseudo-R^2 values with the predictive models for average, fair and olive skin colours were low (0.287, 0.287 and 0.382 respectively) indicating weaker associations.

Of the 17 SNPs reported in various studies to be informative of skin colour (Electronic Supplemental Table S6), six were not included in the 96-plex panel, and another three were removed from analysis. Of the remaining 8 SNPs, only three (rs12896399, rs16891982, rs7495174) were used to model skin colour in this study. However, five SNPs which are known to be strongly associated with eye colour, but not skin colour in particular, were used to build these statistical models: rs1129038, rs1015362 and rs4778138 were used in the dark skin model, rs26722 in both average and dark skin and rs916977 for fair skin (Electronic Supplement Table S6).

There was a lack of concordance between those SNPs reported in the literature as being highly associated with skin colour (Electronic Supplement Table S3), and those used to generate the predictive models in this study (Electronic Supplement Table S6). A published SNP panel for the prediction of skin colour used seven informative SNPs [20]. Of these seven, two (rs6119471 and rs1545397) were not included in this assay, and one (rs885479) was excluded from statistical analysis due to missing data. Of the four remaining SNPs, only two (rs16891982 and rs12203592) were found to be strongly associated with skin colour ($p<0.05$) (Electronic Supplement Table S4). One (rs16891982) of those seven SNPs was used to build the statistical models for all four skin shades. In addition, one of the seven SNPs (rs1426654 in the SLC24A5 gene), reported as highly associated with eye, hair, skin colour and ancestry was not included in any predictive model in this study. The omission of such a high number of SNPs reported to be highly indicative of skin colour from the predictive SNP models generated in this study suggests that the predictive power and accuracy of these phenotype models may be comparatively low. However, the statistical model to predict 'dark skin' versus 'not dark' skin was robust. Dark skin was predicted with 95% accuracy, and 'not dark' skin would be expected to be correct 97% of the time.

Ancestry Prediction

As pigmentation phenotypes vary across population groups and geographical areas, it might be expected that pigmentation SNPs may also be predictive, to some degree, of ancestry. The probability of determining the ancestry of an individual based on the same 96-plex pigmentation SNP panel was investigated. Predictive models were generated for the three major US population groups (Caucasian, African American and Hispanic). The predicted accuracy for all

three population groups was high, with African American being the highest at 96%, Caucasian at 95% and Hispanic at 92%. However, the strength of the association was much lower for the Hispanic model (R^2=0.647 compared to R^2=0.88 in Caucasians) (Fig. 4). This result is supported by the work of Phillips et al. [46] who used a 34-plex SNP assay to predict ancestral origin. All four US population groups (Caucasian, African American, Hispanic and Asian) were reliably predicted, but the classification power for the Hispanic group was much lower than the other three populations tested. This decrease was attributed to the higher degree of admixture within that population.

Because pigmentation SNPs rather than AIMs were selected for the 96-plex SNP assay in this study, only three of the 15 highest ranked SNPs in the Phillips et al. 34-plex [47], were part of our assay (Supplement Table S3). Of the three fixed difference SNPs (rs1426654, rs2814778 and rs16891982) found by Phillips *et al.* to be most informative for ancestry, rs1426654 and rs16891982 were included in the assay tested in this study (Electronic Supplement Table S3). Both SNPs (rs1426654 and rs16891982) were found to be strongly associated with Caucasian ancestry with p-values of 0.005 and 7.21E-08 respectively (Electronic Supplement Table S4). SNP rs1426654 was excluded from the predictive analysis due to >5% missing data, whilst rs16891982 was included in the predictive models for Caucasian and African American ancestry, in addition to all skin colours, all hair colours and all eye colours except green.

Although the accuracy rates for the prediction of ancestry were quite high (>92%), several SNPs found to be strongly associated with ancestry were excluded from the predictive models due to >5% missing data (Electronic Supplement Table S4). The first and second ranked SNPs for African American and Hispanic predictions (1:rs4827380; 2:rs1485682 and 1:rs1805008; 2:rs3212355 respectively) and, the second and third ranked SNPs (rs1800407 and rs1426654 respectively) for the Caucasian group were removed from analyses. If these strongly significant SNPs had been included in models for the prediction of ancestry, even higher accuracies might be expected.

Accuracy of Prediction for each Phenotype

To ascertain the validity of the model predictions for each phenotype, binary logistic regressions were run on an additional 100 blind samples. In 9 of the 15 phenotype categories, the accuracy obtained with blind testing was comparable (within 0.5%) to the accuracy predicted by the statistical model (Table 2). Three phenotype categories (blue, brown and green eyes) performed better than was predicted, and three (hazel eyes, olive skin and Hispanic ancestry) produced slightly lower accuracy results than predicted by the statistical model (Table 2).

Of the four traits investigated in this study, the statistical models for ancestry had the highest accuracy rates for predictions (92.5% – 96.1%). When tested with 100 blind samples, these predictions were accurate for the three population groups. On average the actual accuracy was 0.4% lower than predicted, while the African American predictions were 0.2% more accurate than predicted. The percentage of false negatives and false positives assigned was the lowest in the African American population and the highest with the Hispanic samples. These findings might be expected given that the statistical model for the African American prediction was generally more robust (R2 value = 0.874 compared to 0.647). In addition to the smaller number of Hispanic samples used to predict the statistical model, the Hispanic ethnic group is a heterogeneous population with a high level of admixture, which may also contribute to the higher error rates.

The strength of association of the Caucasian and African American models were similar (R^2=0.885 and 0.874 respectively), and the level of miscalls were moderate (5.4% and 3.7%). However, the actual accuracy of the Caucasian samples was slightly lower than that with the African American samples (94.6% compared to 96.3% respectively).

The predicted accuracy for all four eye colours was similar (83%-86%). The actual accuracy for green and hazel eyes when tested with the blind samples was well predicted by the models (83-87%). The correct call rates for blue eyes and brown eyes were 0.9 and 0.9% more than predicted. This result is expected as blue eye prediction has been reported as high (90-99%) [34, 12, 14]. Although most of these studies have been performed on large datasets of mostly European populations the results of this study support those findings.

The predicted accuracy for skin colour was more variable with a strong prediction for dark skin (97%), 89% for average, and 78-79% for fair and olive skin. The actual accuracy rates when tested were mostly identical (Table 2). The determination of dark skin was the most successful with 96.9%, followed by average skin with 89.8% and both fair and olive skin with approximately 78% accuracy when blind tested.

Blind testing of the hair colour statistical models proved to be moderately accurate. The logistic regression analysis suggested that black and red hair could be accurately predicted ~94% of the time, and blonde and brown colours 88% and 72.8% of the time respectively. The actual accuracies where mostly identical: black hair was as predicted (94.2%), and red (93.2%), blonde (87.6%), and brown (72.9%) hair.

A comparison of GAPIT's mixed linear model with logistic regression showed similar results (see Table 3 for an overview of associated SNPs, three SNPs with the lowest p-value per phenotype. SNPs with a MAF below 0.05 are not listed.)

Comparing the results of the mixed linear model to the logistic regression, rs1129038 has the lowest p-value in both the mixed linear model and logistic regression for the prediction of blue eyes. Another key SNP, rs12913832 has the second lowest p-value in the linear mixed model, but the lowest p-value in the logistic regression for blue and green eyes. After correction for False Discovery Rate, the SNPs with the lowest p-values for green eyes exceeded the threshold of 5%, so the mixed linear model could not generate any valid associations for this trait.

Taken together, the results from the mixed linear model suggest that using a logistic regression in a genome-wide association study is a valid approach as both models share many SNPs in their results. Future GWAS may also benefit from a mix of different methods to confirm that results are as reproducible as possible even when statistical approaches differ.

Conclusion

External visible characteristics such as eye, hair and skin colour are complex polygenetic traits. Many genes and DNA polymorphisms are associated with variations in human pigmentation. As population groups vary in pigmentation phenotypes, pigmentation SNPs may be expected to also infer ancestral origin of a sample. The probability of determining the EVCs and the ancestry of an individual based on the custom 96-plex pigmentation SNP panel was investigated in this research.

In this pilot study, a significant number of samples (~20%) and SNPs (27 out of 96) did not meet the strict requirements for predictive analysis due to poor performance or missing data. Despite this, the SNP panel was able to predict traits such as hair, skin and eye colour in addition to ancestry. In fact, the most robust predictions were obtained when determining ancestry. For the three main US population groups (African American, Caucasian and Hispanic) accurate assignment of ancestry was made 92-95% of the time. The prediction of hair colour was accurate for black and red (94% and 93% accuracy), and reasonably accurate for blonde (88%) and less accurate for brown hair (73%). In general, the ability of this SNP panel to predict eye and skin colour was variable (83-87% and 78-97% respectively). These results were similar to recently published studies [46, 17, 18]. The comparatively low predictive and actual accuracy of some of the traits investigated in this study can be attributed to the fact that several SNPs known to be highly associated with pigmentation were not included in the assay because they were reported in the literature after the panel was designed. Additionally, an unacceptably high number of SNPs and samples were removed from the statistical analysis due to poor performance or missing data.

This SNP panel can provide investigative information regarding the ancestry, eye, skin and hair colour of the donor of a DNA sample, although with variable levels of accuracy. The discriminatory power of this assay could be improved by removal of monomorphic and failing SNPs, and inclusion of the more recently reported SNPs which show strong associations with hair, eye and skin colour. Stronger statistical predictions may also result if a larger sample set from each phenotype was used to generate the predictive models.

Ethical Standards

All appropriate approvals have been granted for this project (Bond University Human Research Ethics Committee approval RO-510) conforming with all Australian ethical requirements.

Acknowledgements

The authors acknowledge the technical assistance provided by Olga Kondrashova and the generous support provided by Bond University Health Sciences & Medicine Faculty, and the National Institute of Justice (NIJ), and the Technical Support Working Group (TSWG) of the United States Government National Interagency Research and Development Program for Combating Terrorism. In addition, substantial in-kind support was provided by Illumina, Inc.

References

1. Budowle B and van Daal A (2008) Forensically relevant SNP classes. Biotechniques 44(5): 603-8, 610.
2. Clark P, Stark AE, Walsh RJ, Jardine R, et al. (1981) A twin study of skin reflectance. Ann Hum Biol 8(6): 529-41.
3. Silventoinen K, Sammalisto S, Perola M, Boomsma DI, et al. (2003) Heritability of adult body height: a comparative study of twin cohorts in eight countries. Twin Res 6(5): 399-408.
4. Graf J, Voisey J, Hughes I, and van Daal A (2007) Promoter polymorphisms in the MATP (SLC45A2) gene are associated with normal human skin color variation. Hum Mutat 28(7): 710-7.
5. Lao O, de Gruijter JM, van Duijn K, Navarro A, et al. (2007) Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann Hum Genet 71(Pt 3): 354-69.
6. Sturm RA (2006) A golden age of human pigmentation genetics. Trends Genet 22(9): 464-8.
7. Voisey J, Box NF, and van Daal A (2001) A polymorphism study of the human Agouti gene and its association with MC1R. Pigm Cell Res 14(4): 264-7.
8. Branicki W, Liu F, van Duijn K, Draus-Barini J, et al. (2011) Model-based prediction of human hair color using DNA variants. Hum Genet 129(4): 443-54.
9. Grimes EA, Noake PJ, Dixon L, and Urquhart A (2001) Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. Forensic Sci Int 122(2-3): 124-9.
10. Rees JL (2000) The melanocortin 1 receptor (MC1R): more than just red hair. Pigment Cell Res 13(3): 135-40.
11. Frudakis T, Thomas M, Gaskin Z, Venkateswarlu K, et al. (2003) Sequences associated with human iris pigmentation. Genetics 165(4): 2071-83.
12. Mengel-From J, Borsting C, Sanchez JJ, Eiberg H, et al. (2010) Human eye colour and HERC2, OCA2 and MATP. Forensic Sci Int Genet 4(5): 323-8.
13. Rebbeck TR, Kanetsky PA, Walker AH, Holmes R, et al. (2002) P gene as an inherited biomarker of human eye color. Cancer Epidemiol Biomarkers Prev 11(8): 782-4.
14. Ruiz Y, Phillips C, Gomez-Tato A, Alvarez-Dios J, et al. (2012) Further development of forensic eye color predictive tests. Forensic Sci Int Genet 7(1): 28-40.
15. Sturm RA and Larsson M (2009) Genetics of human iris colour and patterns. Pigm Cell Melan Res 22(5): 544-62.
16. Walsh S, Liu F, Ballantyne KN, van Oven M, et al. IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet 5(3): 170-80.
17. Walsh S, Liu F, Ballantyne KN, van Oven M, et al. (2011) IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. Forensic Sci Int Genet 5(3): 170-80.
18. Walsh S, Liu F, Wollstein A, Kovatsi L, et al. (2013) The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. Forensic Sci Int Genet 7 (1): 98-115.
19. Liu F, Wen B, and Kayser M (2013) Colorful DNA polymorphisms in humans. Semin Cell Dev Biol.
20. Spichenok O, Budimlija ZM, Mitchell AA, Jenny A, et al. (2010) Prediction of eye and skin color in diverse populations using seven SNPs. Forensic Sci Int Genet 5(5): 472-8.
21. Sturm RA (2009) Molecular genetics of human pigmentation diversity. Hum Mol Genet 18(R1): R9-17.
22. Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, et al. (2010) Predicting phenotype from genotype: normal pigmentation. J Forensic Sci 55(2): 315-22.
23. Bulbul O, Filoglu G, Altuncul H, Freire-Aradas A, et al. (2011) A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. Forensic Sci Int Genet Series 3: e500-501.
24. Butler K, Peck M, Hart J, Schanfield M, et al. (2011) Molecular "eyewitness": Forensic prediction of phenotype and ancestry. Forensic Sci Int Genet Series 3: e498-499.
25. Castel C and Piper A (2011) Development of a SNP multiplex assay for the inference of biogeographical ancestry and pigmentation phenotype. Forensic Sci Int Genet Series 3: e411-412.
26. Kayser M and de Knijff P (2011) Improving human forensics through advances in genetics, genomics and molecular biology. Nat Rev Genet 12(3): 179-92.
27. Relethford JH (1997) Hemispheric difference in human skin color. Am J Phys Anthropol 104(4): 449-57.
28. Investigations FBo. Handbook of Forensic Services. 2007; Available from: http://www.fbi.gov/about-us/lab/handbook-of-forensic-services-pdf.

29. Hosmer DW and Lemeshow S, Applied logistic regression. 2nd ed. Probability and Statistics 2000: Wiley-Interscience Publication.
30. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42(4): 355-60.
31. Lipka AE, Tian F, Wang Q, Peiffer J, et al. (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28(18): 2397-9.
32. Vilhjalmsson BJ and Nordborg M (2013) The nature of confounding in genome-wide association studies. Nat Rev Genet 14(1): 1-2.
33. Keating B, Bansal A, Walsh S, Millman J, et al. (2012) First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. Int J Legal Med: 1-14.
34. Liu F, van Duijn K, Vingerling JR, Hofman A, et al. (2009) Eye color and the prediction of complex phenotypes from genotypes. Curr Biol 19(5): R192-3.
35. Walsh S, Lindenbergh A, Zuniga SB, Sijen T, et al. (2010) Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. Forensic Sci Int Genet 5(5): 464-71.
36. Kastelic V and Drobnic K (2011) Single multiplex system of twelve SNPs: Validation and implementation for association of SNPs with human eye and hair color. Forensic Sci Int: Gen Supp 3: e216-e217.
37. White DL and Rebbeck TR (2012) Using MasterAmp™ PCR PreMixes to Optimize Analysis of High GC Content Genes: PCR Amplification of the Melanocortin-1 Receptor Gene. Epicentre Forum.
38. Pneuman A, Budimlija ZM, Caragine T, Prinz M, et al. (2012) Verification of eye and skin color predictors in various populations. Leg Med (Tokyo) 14(2): 78-83.
39. Visser M, Kayser M, and Palstra RJ (2012) HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. Genome Res 22(3): 446-55.
40. Morgan OW, Sribanditmongkol P, Perera C, Sulasmi Y, et al. (2006) Mass fatality management following the South Asian tsunami disaster: case studies in Thailand, Indonesia, and Sri Lanka. PLoS Med 3(6): e195.
41. Rutty GN, Robinson CE, BouHaidar R, Jeffery AJ, et al. (2007) The role of mobile computed tomography in mass fatality incidents. J Forensic Sci 52(6): 1343-9.
42. Sturm RA (2009) Molecular genetics of human pigmentation diversity. Hum Mol Genet 18(R1): R9-R17.
43. Han J, Kraft P, Nan H, Guo Q, et al. (2008) A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. PLoS Genet 4(5): e1000074.
44. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, et al. (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet 39(12): 1443-52.
45. Gerstenblith MR, Shi J, and Landi MT (2010) Genome-wide association studies of pigmentation and skin cancer: a review and meta-analysis. Pig Cell Melan Res 23(5): 587-606.
46. Phillips C, Fondevila M, Vallone P, Carla S, et al. (2011) Characterization of U.S. population samples using a 34plex ancestry informative SNP multiplex. Forensic Sci Int Genet Series 3: 182-183.
47. Phillips C, Salas A, Sánchez JJ, Fondevila M, et al. (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. Forensic Sci Int: Genetics 1(3-4): 273-280.

Figure 1

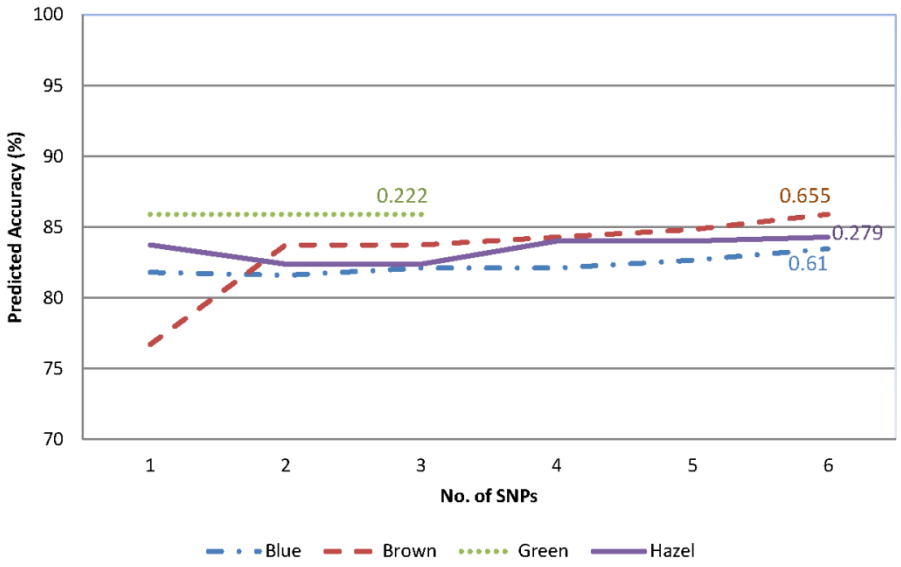


Fig. 1 Comparison of the statistical models used to predict eye colour

The predicted accuracy of each statistical model for each eye colour is compared. Values displayed are Nagelkerke R^2 values for each model indicating the goodness-of-fit.

Figure 2

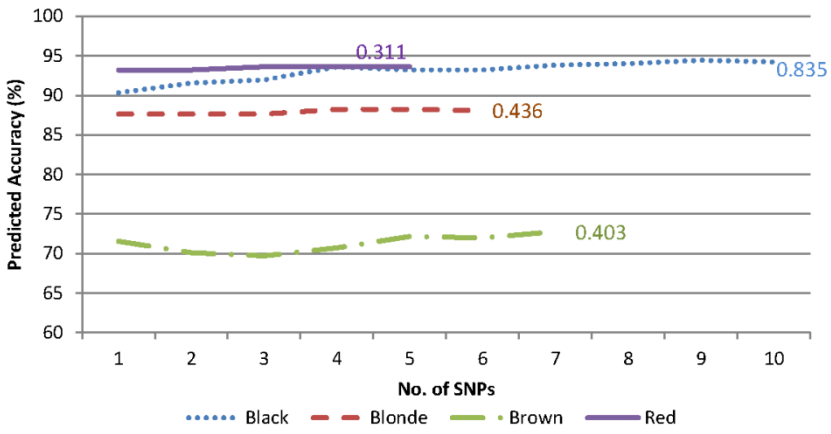


Fig. 1 Comparison of the statistical models for the prediction of hair colour

The predicted accuracy of each statistical model for each hair colour is compared. Values displayed are Nagelkerke R^2 values for each model indicating the goodness-of-fit.

Figure 3

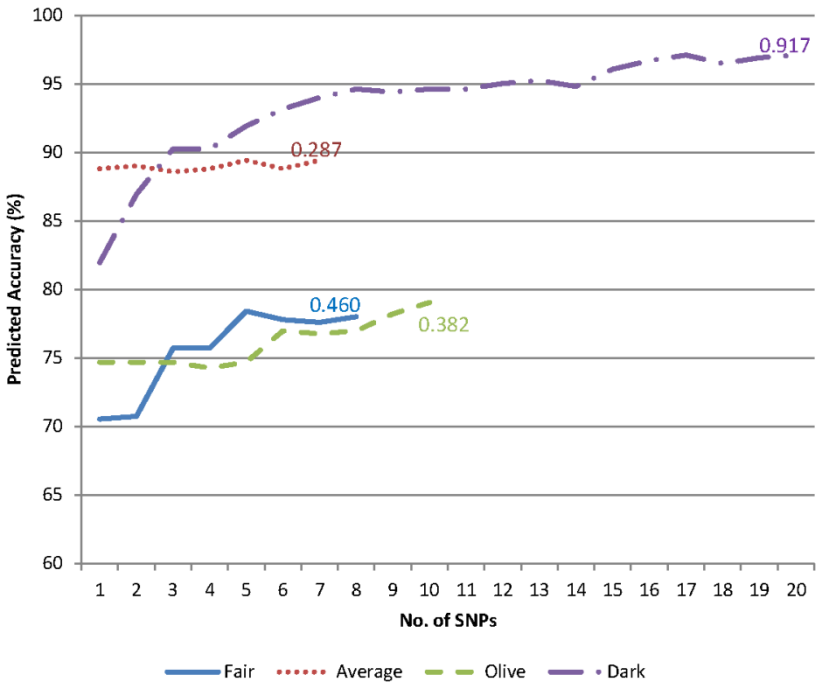


Fig. 1 Comparison of the statistical models for the prediction of skin colour

The predicted accuracy of each statistical model for each skin colour is compared. Values displayed are Nagelkerke R^2 values for each model indicating the goodness-of-fit.

Figure 4

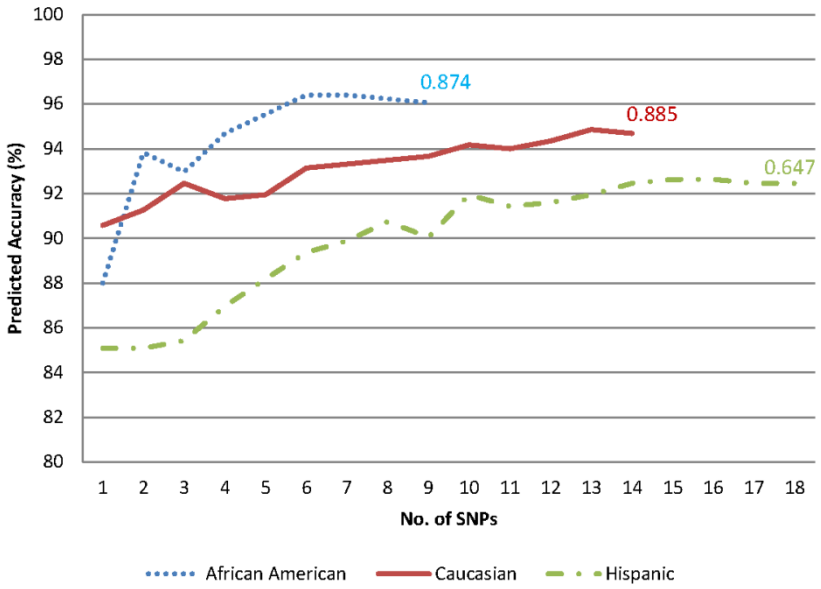


Fig. 1 Comparison of the statistical models for the prediction of ancestry

The predicted accuracy of each statistical model for each population group is compared. Values displayed are Nagelkerke R^2 values for each model indicating the goodness-of-fit.

Table 1

Table 1. The number of DNA samples genotyped and subsequently used for the generation of predictive models for each phenotype

		No. Samples			
	Trait	**[a]Complete Dataset**	**[b]Available Dataset**	**[c]Used for Prediction model**	**Used for Prediction model (%)**
Ethnicity	African American	181	160	119	74.38
	Caucasian	453	371	298	80.32
	Hispanic	167	140	96	68.57
Skin	Fair	215	175	142	81.14
	Average	177	147	122	82.99
	Olive	80	67	57	85.07
	Dark	237	210	161	76.67
Eyes	Blue	176	146	115	78.77
	Brown	205	167	142	85.03
	Green	77	60	52	86.67
	Hazel	88	75	60	80.00
Hair	Blonde	96	76	60	78.95
	Brown	287	233	190	81.55
	Red	45	38	33	86.84
	Black	298	255	202	79.22

[a] Total number of samples genotyped

[b] No. of samples available for predicting models after 100 random samples were removed

[c] No. of samples used to generate predictive models

Table 2. Blind testing of the predictive models for each phenotype.

Trait	Phenotype	Predicted Accuracy (%)	Actual Accuracy (%)	False Negatives (%)	False Positives (%)
Ancestry	Caucasian	94.7	94.6	2.54	2.82
	African American	96.1	96.3	2.46	1.23
	Hispanic	92.5	91.7	5.33	2.95
Hair Colour	Black	94.2	94.27	3.47	2.26
	Blonde	88.0	87.6	8.08	4.29
	Brown	72.8	72.9	11.38	15.76
	Red	93.6	93.24	5.74	1.01
Skin Colour	Fair	78.0	77.97	11.36	10.84
	Average	89.4	89.8	8.88	1.34
	Olive	79.0	77.8	16.49	5.67
	Dark	97.1	96.9	1.74	1.39
Eye Colour	Blue	83.5	84.4	5.73	9.86
	Brown	85.9	86.8	8.88	4.33
	Hazel	84.3	83.4	14.25	2.3
	Green	85.9	86.5	13.51	0.0

* n=100 individuals

Table 3. Associated SNPs as reported by the mixed linear model

Trait	Phenotype	Associated SNP	p-value	FDR-adjusted p-value	Minor Allele Frequency
Ancestry	Caucasian	rs16891982	7.2e-8	6.91e-6	.401
	Caucasian	rs1800407	1.39e-4	.006	.09
	Caucasian	rs1426654	.004	.15	.34
	African American	rs4827380	2.2e-5	.002	.25
	African American	rs1485682	.0005	.026	.24
	African American	rs1998076	.0011	.034	.39
	Hispanic	rs1805008	1.13e-5	.001	.33
	Hispanic	rs3212355	8.03e-5	.003	.33
	Hispanic	rs35264875	.0001	.004	.14
Hair Colour	Black	rs16891982	1.25e-7	1.20e-5	0.36
	Black	rs3212355	.003	.09	.35
	Black	rs1667394	.01	.25	.38
	Blonde	rs1110400	.23	1	.48
	Blonde	rs12203592	.344	1	.15
	Blonde	rs1129038	.41	1	.48
	Brown	rs16891982	.13	1	.39
	Brown	rs1129038	.14	1	.48
	Brown	rs1805005	.23	1	.09
	Red	rs4911442	.14	1	.09
	Red	rs1805005	.45	1	.09
	Red	rs12821256	.48	1	.08
Skin Colour	Fair	rs16891982	.02	.99	.36
	Fair	rs885479	.1	.99	.21
	Fair	rs1129038	.13	.99	.48
	Average	rs1900758	.02	.86	.48
	Average	rs6137444	.03	.86	.44
	Average	rs12203592	.05	.86	.14
	Olive	rs1805005	.011	.65	.09
	Olive	rs16891982	.013	.65	.36
	Olive	rs12211228	.02	.69	.13
	Dark	rs16891982	4.4e-8	4.26e-6	.36
	Dark	rs7495174	.001	.07	.13
	Dark	rs26722	.003	.08	.07
Eye Colour	Blue	rs1129038	3.2e-9	3.15e-7	.38
	Blue	rs12913832	.0001	.007	.4
	Blue	rs1110400	.06	1	.48
	Brown	rs1129038	7.4e-7	7.1e-5	.38
	Brown	rs12913832	3.2e-5	.001	.4
	Brown	rs16891982	.0008	.02	.23
	Hazel	rs1129038	2.63e-8	2.53e-6	.38
	Hazel	rs12913832	5.23e-7	2.51e-5	.401
	Hazel	rs1426654	.01	.24	.204
	Green	rs12896399	.001	.11	.42
	Green	rs12913832	.003	.16	.4
	Green	rs1129038	.008	.27	.38

3.6. Ion Torrent platform evaluation and optimization

3.6.1. Introduction

The massively parallel sequencing (MPS) on the Ion Torrent platform provides unprecedented coverage of thousands of markers using only 10 ng of template (or less with additional amplification cycles). At the same time, it is relatively novel and is not yet a routine method (as discussed in section 1.10). It is prone to fluctuations in genomic DNA input, library and template preparation efficiency (e.g. frequent instruments malfunction or DNA amplification and/or enrichment efficiency); chip loading (e.g. the angle of pipet's tip while loading the sample and the way of centrifuging the chip) and data analysis algorithms, which are frequently updated. As a result, the replicates of the same sample on different chips cannot be compared accurately by their output (total number of SNPs) or sequencing depth, but only by the accuracy of genotyping (concordance between replicates).

The overarching goal of this study was to undertake a pilot study, evaluating the Ion Torrent platform for subsequent genotyping of hundreds of DNA samples and pinpoint the steps that could be optimised for improved method performance.

The main aims of this study were:

a) To determine the effect of half (10µl) versus full (20µl) volume of the library amplification reaction.

   The manufacturer's protocol recommends using a full 20µl volume per library amplification reaction. However, given that custom primers were designed in two multiplex pools, this required using double the amount of sequencing consumables, significantly increasing the cost and processing time. A half reaction volume is mentioned as an option in the protocol, although not officially supported by the manufacturer [340].

b) To optimize the library concentration input per OneTouch amplification and enrichment (10pM vs. 20pM).

   An accurate estimation of library concentration used for template amplification and enrichment is important, as it highly correlates with sequencing output, although it varies significantly among PGM instruments and sequencing kits. The manufacturer recommends using between 10pM and 20pM input, while even a slight difference (e.g. 2pM) can make a difference. However, some

laboratories use between 8pM to 25pM DNA concentration, based on internal validation (personal communication with other laboratories and web forums).

c) To validate and optimise sequencing throughput (maximum number of DNA samples) that can be genotyped on a 316 chip at a minimum depth of x20.

   The manufacturer supported output of the 316 chip is 6.2 million reads and between 30 Mb to 200 Mb per run. However, it has been shown that this output can be significantly increased, mostly by using in-house chip-loading methods (as discussed on the Ion Torrent community web server – http://ioncommunity.lifetechnologies.com/welcome).

d) To confirm genotype concordance between sample duplicates.

   The genotyping accuracy is the most important criterion of the sequencing, especially in the forensic DNA context. Comparing the sequencing output of the DNA sample duplicates or preferably triplicates is an important part of the concordance check.

It should be noted that due to the high cost of the sequencing consumables, only a limited number of biological duplicates were assessed in this study.

3.6.2. Materials and Methods

The following items in this section (items a to d) correspond to the aims listed in the Introduction section 3.6.1.

a) Three DNA samples at an initial concentration of 10ng per reaction were used to validate the full (20µl) versus half (10µl) Ampliseq library amplification reaction. The samples with half reaction volume were merged after the first amplification step, as per manufacturer recommendations (Chapter 2.9.1). Specific Ion Xpress barcode X (X= the number of the specific barcode) adapters (Life Technologies) were ligated onto the 5′ and 3′ ends of each fragment and linked by nick translation (as per manufacturer recommendations) to allow for sequencing multiple samples simultaneously (Chapter 2.9.2).

b) Ampliseq libraries were prepared for 32 DNA samples in half reaction volume (10µl) in duplicates. Library quantities were determined using the Ion Library Quantitation

kit (Life Technologies). An input of 10pM and 20pM (per each set of 32 samples) was used for amplification by emulsion PCR using the OneTouch instrument (Life Technologies). The libraries were subsequently enriched using the OneTouch ES instrument (Life Technologies).

c) DNA samples (n=6, 16 or 32) were sequenced on three 316 chips. The sequencing was performed on the PGM (Life Technologies) according to the manufacturer recommendations (Chapter 2.9.4). The sequencing output was assessed using the overall chip output (aligned reads), mean sequencing depths and number of targeted markers per samples.

d) Biological duplicates (n=38) and triplicates (n=8) were genotyped by sequencing and analysed using the Ion Torrent suite software with the built-in Variant Caller (VC) plugin (version 3.2) as well as a cloud-based Ion Reporter software (versions 1.2, 1.4 and 1.6). The VC parameters were as follows: library type, Ampliseq; targeted region, custom designed targets list; hotspot list, all SNPs in the dbSNP132 database. Variants were detected using the low stringency germ line parameters. The rest of the parameters were as per default. The VC output was present in tabular format (.vcf file), as a list of differences between the samples and a reference sequence (hg19). The ".vcf" files for all the relevant samples were uploaded into the SVS software package (GoldenHelix) and analysed for concordance in duplicates or triplicates.

3.6.3. Results and Discussion

Of the 587 DNA samples genotyped, the majority were sequenced at high coverage (at least x20) and only three samples produced no results.

A validation of the full versus half amplification reaction volume showed that sample outputs of the half reactions performed as well as the full volume reactions (Table 7). The overall chip output (for both volumes) was significantly better, compared to the value officially supported by the manufacturer – 410 Mb versus 200 Mb. The overall chip output is a function of various parameters, although the chip loading is the most important.

In spite of the limited number of samples tested (only three duplicates sequenced on one chip), the results demonstrated that the use of a half volume reaction produced the same quality output as the full volume reaction. These conclusions were supported by the output of subsequent chips (n=35) that were processed.

Table 7. A summary of the chip output for half versus full reaction volume experiment.

Barcode #	Sample name	Bases	>=Q20 bases	Reads	Mean read length	Read Depth	Variants detected
IonXpress_001	10-001 full volume	103,192,101	87,793,917	716,854	143 bp	322.05	1,070
IonXpress_002	10-005 full volume	39,448,588	33,845,004	269,476	146 bp	131.51	1,084
IonXpress_003	10-007 full volume	64,274,573	55,419,322	431,752	148 bp	226.59	1,019
IonXpress_004	10-001 half volume	74,266,749	63,486,418	515,672	144 bp	230.98	1,070
IonXpress_005	10-005 half volume	63,038,639	53,912,744	432,493	145 bp	210.47	1,067
IonXpress_006	10-007 half volume	66,704,358	57,550,238	449,074	148 bp	235.11	1,010

The validation of the library concentration input per OneTouch reaction showed that samples at 10pM concentration performed better than at 20pM (Tables 8 and 9). The number of polymorphisms detected in the samples at 10pM were slightly higher than in the same samples at 20pM (5% difference). Polyclonality was the most significantly affected parameter of the run (16% at the 10pM versus 27.4% at the 20pM). Polyclonality is the number of ISPs that have more than a single targeted clone and as a result, cannot be used for sequencing and alignment. Generally, the higher the concentration of enriched beads, the higher the polyclonality was, which may affect the sequencing output. The sequencing depth of the 10pM chip was between 73.82 (maximum) and 4.12 (minimum) compared to 58.67 (maximum) and 6.29 (minimum) of the 20pM chip. Following these results, the 10pM library concentration was found optimal for all the subsequent samples.

Table 8. A summary of 10pM input concentration per chip experiment.

	AQ17	AQ20	Perfect		
Total Number of Bases [Mbp]	419.47	363.39	282.68		
Mean Length [bp]	141	127	101		
Longest Alignment [bp]	313	303	296		
Total reads	3,117,857				
Usable reads	77%				
Addressable Wells	**6,348,217**		**Library ISPs**	**4,058,401**	
With ISPs	4,156,532	65.50%	Filtered: Polyclonal	647,820	16.00%
Live	4,147,423	99.80%	Filtered: Low Quality	283,122	7.00%
Test Fragment	89,022	2.10%	Filtered: Primer Dimer	9,602	0.20%
Library	4,058,401	97.90%	**Final Library ISPs**	**3,117,857**	**76.80%**

Average value of 10pM input				
Mapped Reads	On Target	Mean Depth	Uniformity	Variants detected
96,386	90.20%	46.6	89.80%	972

Table 9. A summary of the 20pM input concentration per chip experiment.

	AQ17	AQ20	Perfect		
Total Number of Bases [Mbp]	332.44	271.35	214.72		
Mean Length [bp]	118	102	83		
Longest Alignment [bp]	297	283	275		
Total Reads	3,151,433				
Usable reads	67%				
Addressable Wells	**6,348,217**		**Library ISPs**	**4,689,035**	
With ISPs	4,856,455	76.50%	Filtered: Polyclonal	1,285,637	27.40%
Live	4,738,662	97.60%	Filtered: Low Quality	244,263	5.20%
Test Fragment	49,627	1.00%	Filtered: Primer Dimer	7,702	0.20%
Library	4,689,035	99.00%	**Final Library ISPs**	**3,151,433**	**67.20%**

Average value of 20pM input				
Mapped Reads	On Target	Mean Depth	Uniformity	Variants detected
89,031	91.38%	37.91125	89.57%	927.90625

Sequencing of 6, 16 and 32 samples per 316 chip demonstrated that chip output (total bases) can be significantly increased above the manufacturer supported 200 Mb to more than 600 Mb. The increased output was also a function of ISP performance, which was optimal at approximately 10pM concentration. The mean sequencing depth was lower when a higher number of samples per chip was used (n=32), albeit sufficient for comprehensive coverage of targeted markers.

Alteration to the original chip loading protocol were assessed as recommended on the Ion Torrent community forum (http://ioncommunity.lifetechnologies.com/) and also following in-house changes (performed by the author), had the greatest influence on increased chip output. Various changes included increasing the number of isopropanol and annealing buffer washes (three versus one in the protocol); an additional centrifugation step and vortex mixing of the loaded chip; blowing the chip with argon gas after the washing step and careful loading of the ISPs to prevent potential air bubbles from appearing. Based on the results of this study, up to 32 barcoded samples were processed on each 316 chip.

The concordance study of 38 DNA duplicates and eight (8) triplicates, showed various discrepancies in allele calls. The discrepancies between duplicates usually included only one nucleotide (one DNA strand) difference in SNP (e.g. AC vs. AA or TT vs. GT) and a lack/addition of one repeat in INDEL (-/GG vs. GG_GG or TCTAG_TCTAG vs. -/TCTAG). A missing call in one of the replicates was not considered a discrepancy. The two markers that most frequently showed inconsistency in allele calls were the chromosome 4:81975674 and chromosome 6:137345858, both INDELs, which are more prone to sequencing errors.

The overall discrepancy among samples analysed at low stringency variant detection settings, varied between 2 to 19 markers per sample pair corresponding to 0.29% to 1.9% of the total number of polymorphisms. In general, samples sequenced at a higher depth showed less discrepancy (three to four markers per sample pair or approximately 0.3%), as per Table 10. However, the percentage of inconsistent calls might not be an accurate representation, since the total number of markers can be counted in various ways, such as the total number of genotyped markers or successfully genotyped markers only.

Table 10. Example of three samples duplicates, sequenced on the same chip and analysed under low stringency algorithm. The discrepancy in allele calls and its percentage from the total number of calls is shown.

		sequencing depth		
Sample ID	Chromosome location	322.05	230.98	percentage from the total number of markers
10-001	2:223056368-SNV	A_C	C_C	0.37
	6:137345858-Ins	AA_AA	AA_TAA	
	7:130713701-Del	-_C	A_C	
	11:2192316-Ins	TGAA_TGAA	-_TGAA	
		sequencing depth		
		131.51	210.47	
10-005	1:242342504-SNV	C_T	T_T	0.37
	2:152813963-Del	-_GG	GG_GG	
	4:179555381-SNV	A_T	T_T	
	10:131092508-Ins	-_GGAA	GGAA_GGAA	
		sequencing depth		
		226.59	235.11	
10-007	4:42003383-SNV	C_T	C_C	0.29
	6:137345858-Ins	AA_TAA	AA_AA	
	13:31803958-Del	G_T	-_-	

The output for the same samples, compared in triplicates with various sequencing depth is summarized in Table 11.

Table 11. Comparison between triplicates of three samples, sequenced on two chips and analysed using low stringency parameters.

		sequencing depth			percentage of total markers
Sample	Chr. location	322.05	230.98	40.38	
10-001	2:223056368-SNV	A_C	C_C	?_?	1.13
	4:42003671-Mix	G_G	G_G	A_G	
	4:155508898-Del	-_AGAAAGAA	-_-	?_?	
	5:112568212-SNV	T_T	T_T	C_T	
	6:21911616-SNV	G_G	G_G	G_T	
	6:137345858-Ins	AA_AA	AA_TAA	?_?	
	7:132192020-SNV	G_G	G_G	A_G	
	8:61606762-Del	C_C	C_C	-_C	
	8:61726879-SNV	A_G	A_G	A_A	
	8:139816051-Del	G_T	G_T	-_G	
	10:17025666-SNV	C_C	C_C	C_T	
	11:2192316-Ins	TGAA_TGAA	-_TGAA	?_?	

	14:55585558-Sub	C_T	C_T	C_CAT	
	15:33023670-SNV	T_T	T_T	C_T	
	16:7587676-SNV	T_T	T_T	G_T	
	22:19759437-SNV	C_G	C_G	C_C	
	X:13779124-SNV	A_T	A_T	A_A	
		sequencing depth			percentage of total markers
Sample	Chr. location	131.51	210.47	11.18	
10-005	1:103412000-SNV	C_C	C_C	C_T	1.38
	1:242342504-SNV	C_T	T_T	?_?	
	1:242806860-Mix	T_T	T_T	A_T	
	2:16767106-SNV	G_G	G_G	A_G	
	2:42577804-Ins	AAATACACAC_AAATACACAC	AAATACACAC_AAATACACAC	-_AAATACACAC	
	2:152813963-Del	-_GG	GG_GG	?_?	
	2:224794578-Ins	AG_AG	AG_AG	-_AG	
	4:81975674-SNV	A_G	?_?	G_G	
	4:179555381-SNV	A_T	T_T	?_?	
	5:89818764-SNV	G_G	G_G	A_G	
	5:123111246-SNV	A_A	A_A	A_C	
	5:174157972-SNV	C_C	C_C	C_G	
	7:132192020-Del	A_G	A_G	-_A	
	8:61595671-SNV	C_C	C_C	C_T	
	8:61652144-SNV	A_A	A_A	A_G	
	10:115316725-SNV	A_C	A_C	A_A	
	11:18317458-SNV	A_A	A_A	A_T	
	14:90077221-SNV	C_C	C_C	C_T	
	18:19651889-Mix	C_C	C_C	A_C	
	22:26862212-Del	C_T	C_T	-_T	
		sequencing depth			percentage of total markers
Sample	Chr. location	226.59	235.11	20.81	
10-007	1:103568727-SNV	A_T	A_T	T_T	1.06
	2:109401292-Ins	TCTAG_TCTAG	TCTAG_TCTAG	-_TCTAG	
	2:223055226-SNV	?_?	A_C	C_C	
	4:42003383-SNV	C_T	C_C	C_T	
	4:81975674-SNV	?_?	A_G	G_G	
	6:137345858-Ins	AA_TAA	AA_AA	AA_AA	
	13:31803958-Del	G_T	-_-	?_?	
	20:25278465-Ins	GTGGG_GTGGG	GTGGG_GTGGG	-_GTGGG	
	20:55765485-SNV	A_A	A_A	A_G	
	21:17710424-SNV	A_A	A_A	A_G	
	22:46808261-SNV	G_G	G_G	C_G	

The inconsistency in allele calls observed in replicates is most likely related to the parameters of sequence alignment algorithms (e.g. low/high stringency) and coverage. Three duplicates (see Table 10) sequenced at significantly higher depth, showed less discrepancy between each other, compared to the same samples, sequenced at approximately ten fold less depth. In addition, the same duplicates while analysed according to the high stringency parameters, showed significantly higher concordance (from zero to only three markers difference) per duplicate (Table 12). On the other hand, applying the higher stringency settings, resulted in less markers detected (down by approximately 30%).

Table 12. The same samples as in Table 10, analysed using Variant Caller, according to high stringency parameters.

Sample ID	Chromosomal location	Variant		% of the total number of markers
10-001	7:130713701	A_C	-_C	0.096
10-005	4:42003360	A_G	G_G	0.287
	4:42003383	C_T	C_C	
	4:179555381	A_T	T_T	
10-007	no discrepancies			

The main differences between the low stringency and the high stringency analysis settings are summarized in Table 13.

Table 13. Comparison between the low and high stringency settings in the Variant Caller plugin.

	Germ line - Low stringency			Germ line - High stringency		
General parameters	SNP	INDEL	Hotspot	SNP	INDEL	Hotspot
min cov each strand	0	5	3	3	3	3
min variant score	10	10	10	10	10	10
min allele freq	0.1	0.1	0.1	0.15	0.15	0.15
min coverage	6	15	6	20	20	20
strand bias	0.95	0.85	0.95	0.95	0.85	0.95
Torrent Variant Caller parameters	**Germ line - Low stringency**			**Germ line - High stringency**		
data_quality_stringency	6.5			8.5		
hp_max_length	8			8		
filter_unusual_predictions	0.3			0.25		
filter_insertion_predictions	0.2			0.2		
filter_deletion_predictions	0.2			0.2		
snp_beta_bias	30			8		
indel_beta_bias	8			8		
hotspot_beta_bias	30			8		
downsample_to_coverage	400			400		
outlier_probability	0.01			0.01		
do_snp_realignment	0			0		
prediction_precision	1			1		
heavy_tailed	3			3		
suppress_recalibration	1			1		
Long Indel Assembly Settings						
kmer_len	19			19		
min_var_count	5			5		
short_suffix_match	5			5		
min_indel_size	4			4		
max_hp_length	8			8		
min_var_freq	0.15			0.15		
relative_strand_bias	0.8			0.8		
FreeBayes						
allow_indels	1			1		
allow_snps	1			1		
allow_mnps	0			0		
min_mapping_qv	4			4		
min_base_qv	4			4		
read_mismatch_limit	10			10		
read_max_mismatch_fraction	1			1		
gen_min_alt_allele_freq	0.15			0.15		
gen_min_indel_alt_allele_freq	0.15			0.15		
gen_min_coverage	6			6		

For the purpose of the association test, sequences were analysed using the Ion Reporter software with at least x6 depth threshold per SNP. However, prior to statistical analysis (Chapter 4) the data were filtered by a minimum coverage of x10 and genotype quality of 10 (per marker), thus reducing the potential error rate.

The demonstrated lack of concordance of up to 1.9% (at low stringency parameters) may be adequate for research purposes, although it is unacceptable for forensic DNA analysis. It should be noted, that approximately 20% of sequencing errors observed in this study were within INDEL variations, which are more prone to sequencing errors due to homopolymer repeats. Remarkably, an application of the high stringency analysis parameters resulted in less than 0.3% discrepancy in allele calls between samples. Therefore, for any forensic use, a high coverage (x100 or higher) followed by high stringency variant analysis parameters should be considered. This approach would allow significantly more accurate sequencing (as the first priority in forensic DNA analysis), although significantly increasing cost and decreasing throughput on the other hand.

Most of the currently available MPS platforms demonstrate high variability in sequencing accuracy and are not yet optimal in terms of hardware and software updates. For example, the Ion Torrent hardware, software, reagents and laboratory manuals in this project were updated through series of optimization as follows:

- The OneTouch instrument (an essential component for template preparation) was physically replaced five times in one year and its software was therefore also updated five times.
- Template preparation and sequencing kits were updated and changed three times.
- The PGM software and the Ion Torrent suite software were updated ten times in one year from version 2.1 to 3.6.
- The manuals for library preparation, template preparation and sequencing were updated four times during the project.
- The Ion Reporter software used for data analysis was updated four times (from version 1.2 to 4).

The update of one platform's component (either software, hardware or reagents) led to a subsequent adaptation of all other parts of the Ion Torrent platform, which usually were irreversible and non-compatible with previous versions. A more standardized procedure

is required in order to increase the robustness of the Ion Torrent platform, prior to its potential use for forensic DNA typing. However, given the dynamic nature of the MPS technologies, these problems are likely to be addressed in the near future.

A limited validation performed in this study tested the concordance and genotyping accuracy within the same platform. Given the relatively novel nature of the Ion Torrent platform, it was decided to perform a limited comparison between at least two independent genotyping methods. Due to the fact that a subset of DNA samples (n=71) were genotyped for 78 SNPs that overlapped between the Golden Gate assay and the Ion Torrent platform, this comparison was possible. However, due to the different focus and time limits of this project, this validation was not meant to be extensive, and included only a subset of samples and SNPs. A limited comparison of five (5) SNPs in all samples genotyped using both platforms, was performed. The SNPs that were analysed included rs12913832, rs1129038, rs611349, rs739289 and rs12203592. The analysis demonstrated only one inconsistency between genotypes of the respective samples ('TC' versus 'CC' in the rs12203592). Given that these two platforms use completely different methods for genotyping, this outcome can be considered a successful quality control step.

In order to properly test the concordance and genotyping accuracy within and between platforms, a larger set of DNA samples and markers should be genotyped by both platforms. However, the results of this preliminary study provided important verification of the genotyping output obtained within the Ion Torrent platform, as detailed in this section. It should be noted that additional verification of the Ion Torrent genotyping output was obtained by performing association analyses of the pigmentation traits and ancestry (as discussed in Chapter 5). The results obtained in this study were in full concordance with previously published results.

Chapter 4

Optimization of scanning equipment and 3D image processing

4.1. Introduction

The main goal of this study was to assess the craniofacial measurements reproducibility and normal distribution, as per Aims 2 and 3 of the current project (Section 1.10).

Prior to investing in the 3D technology, the validity of the anthropometric measurements using 2D images was tested. This study was required because the 3D scanner was not available at the beginning of this project and since it is generally easier to generate 2D photographs, which can be taken at various locations without the need to attend a scanning session in a dedicated room, as required for the 3D scanning approach.

Locating anatomical landmarks on the face and obtaining anthropometrical measurements from photographs is known as photogrammetry [17]. This method has been tested in numerous anthropometrical studies [29, 36, 351-353]. This process is easier and less time consuming than the direct measurements and 3D digital image processing. On the other hand, it possesses significant disadvantages, such as the quality of photographs and thus resulting in a partial coverage of facial landmarks. Although most importantly, it is unable to reflect the actual surface distances, but can only appreciate the lateral facial distances. The lateral measurement is the linear distance between two landmarks, which does not necessarily represent the actual physical (Euclidean) distance over the skin surface. As a result, facial measurements obtained from 2D images are less accurate than direct or 3D–acquired measurements, hence providing only partial and potentially misleading information on facial morphology.

Conversely, 3D scanning platforms provide a more comprehensive representation of the facial morphology. The 3D scanning systems have been extensively used in anthropometric studies as well as in medical research, which usually demands a high accuracy and precision [354-357]. The Minolta Vivid V910 3D scanner that was acquired for this study, has a reported manufacturing precision of ± 0.008mm and accuracy of ±0.3 mm. A validation study showed that it is accurate to a level of 1.1±0.3mm [358], while another study demonstrated a level of accuracy of 0.56± 0.25mm and the error in computerized registration of left and right scans as 0.13± 0.18 mm [359]. These observations demonstrate that this platform is more appropriate for the present study and would be expected to provide a more accurate representation of the facial morphology as compared to less accurate 2D photographs.

This study also included a comparison of lateral and surface measurements generated from 3D images. The rationale behind this comparison was that it is easier and faster to generate the lateral, rather than the surface measurements in the Geomagic software.

4.2. Materials and Methods

Facial measurements of ten individuals, obtained from lateral dimensions on 2D images and lateral and surface measurements generated from 3D images were compared. The 2D photographs (n=10) were taken using a Cyber-shot DSC-T100 digital camera (Sony) with an approximate resolution of 10 megapixels. The 2D photographs were taken under ambient lighting, similar to the 3D scanning settings. The volunteers sat holding a ruler near their faces. The photographs were printed on A4 paper and analysed for a set of 13 facial measurements. Several measurements were not performed due to image quality (as discussed in Section 4.3).

The scanner used in this study was sensitive to changes in light conditions and was calibrated every three months (according to manufactorer's recommendations) to produce high quality standardised 3D images. The 3D facial scans (n=10) and image processing were performed as described in the Materials and Methods Chapter 2.

The quality of the 3D alignment was measured automatically by Polygon software and displayed as an "error average" and "sigma" with values between 0 and 1 (illustrated in Figure 27). The final image was considered of good quality if both values were below 0.4 (as per manufacturer recommendations). If the values were above 0.5 the volunteer was re-scanned. The measurements obtained from 2D and 3D images of ten (10) individuals were recorded in an Excel spreadsheet and compared manually.

4.3. Results and Discussion

The 2D photogrammetry provided easier image processing and used inexpensive equipment, compared to 3D scanning. The 3D scanning not surprisingly, was found to be significantly more accurate and comprehensive. However, the Geomagic software (used for 3D image processing), was not designed for craniofacial measurements and

elicited significant difficulties. The need for measuring the Euclidian distance between landmarks, rather than the lateral distance required a special assistance from the manufacturer, although this support was only partial and time consuming. In addition, each of the 32 measurements were performed and recorded individually, which made the landmarking procedure tedious and time-consuming, requiring approximately 20 minutes per image. Attempts to generate a 'macro' function, which would automatically copy a set of specific facial landmarks to a new image using Geomagic program, failed, as each face was significantly different from the previous and the software could not perform this procedure in an efficient manner.

Comparison of the measurements showed that lateral and surface measurements performed on the 3D digital images were noticeably different. The 3D surface distances were longer than the lateral, with the latter more similar to the 2D measurements. The results of the comparison between 2D and 3D measurements are summarised in Tables 14, 15 and Figures 42, 43.

Since it was obvious that there was a pronounce difference and given that 3D surface measurements represent the most adequate information on facial features, it was concluded that all the measurements should be calculated using Euclidean coordinates of the craniofacial landmarks (representing the actual surface distance), which was performed using Microsoft Excel automatic spreadsheet. The 3D surface measurements were subsequently used as phenotypes for genetic association study, as detailed in Chapter 5.

Table 14. Results of the comparison between craniofacial measurements in 2D and 3D images, including lateral and surface distance. The values shown are in millimetres.

	Volunteers														
	1			2			3			4			5		
	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo
n-gn	123.41	142.88	124.00	114.75	138.88	117.00	113.48	134.20	115.00	137.74	171.33	135.00	114.46	141.02	113.00
zy-zy	135.53	174.02	135.00	124.73	171.11	121.00	123.47	168.02	128.00	135.72	189.97	127.00	137.69	189.30	127.00
go-go	111.55	137.12	116.00	114.35	156.48	107.00	103.15	132.28	114.00	108.41	141.94	103.00	123.36	161.33	113.00
en-en	46.15	65.65	38.00	34.55	57.49	31.00	33.52	52.94	33.00	34.96	54.75	32.00	36.20	56.62	30.00
ec-ec	114.25	135.51	106.00	94.15	126.84	90.00	97.14	124.71	97.00	95.61	119.15	93.00	95.30	119.72	91.00
n-sn	50.19	61.66	51.00	49.86	57.73	55.00	46.08	59.49	48.00	55.74	66.31	52.00	53.00	65.04	54.00
al-al	38.40	48.26	39.00	38.42	50.23	38.00	37.86	51.26	38.00	35.43	51.64	39.00	36.49	52.67	43.00
ls-li	19.77	23.55	18.00	15.36	19.71	13.00	18.23	19.61	18.00	23.54	26.08	25.00	17.59	19.77	17.00
ch-ch	50.00	56.34	50.00	47.26	55.47	48.00	48.04	51.16	55.00	49.75	59.20	55.00	52.42	62.20	53.00
sn-sto	27.67	29.37	26.00	24.57	28.32	21.00	20.77	22.83	20.00	28.17	30.20	26.00	16.53	20.56	15.00
sn-gn	76.06	84.23	73.00	66.82	79.16	63.00	69.09	79.74	66.00	88.62	102.36	83.00	64.14	80.21	59.00
sto-lm	21.87	25.38	21.00	15.97	18.56	13.00	20.26	23.30	18.00	19.55	22.53	26.00	19.81	24.16	16.00
left ear sa-sba	61.85	72.97	66.00	NA	NA	62.00	58.70	70.54	58.00	62.17	70.57	62.00	NA	NA	71.00
right ear sa-sba	65.29	72.96	72.00	NA	NA	63.00	55.36	76.51	53.00	61.26	79.53	63.00	NA	NA	72.00

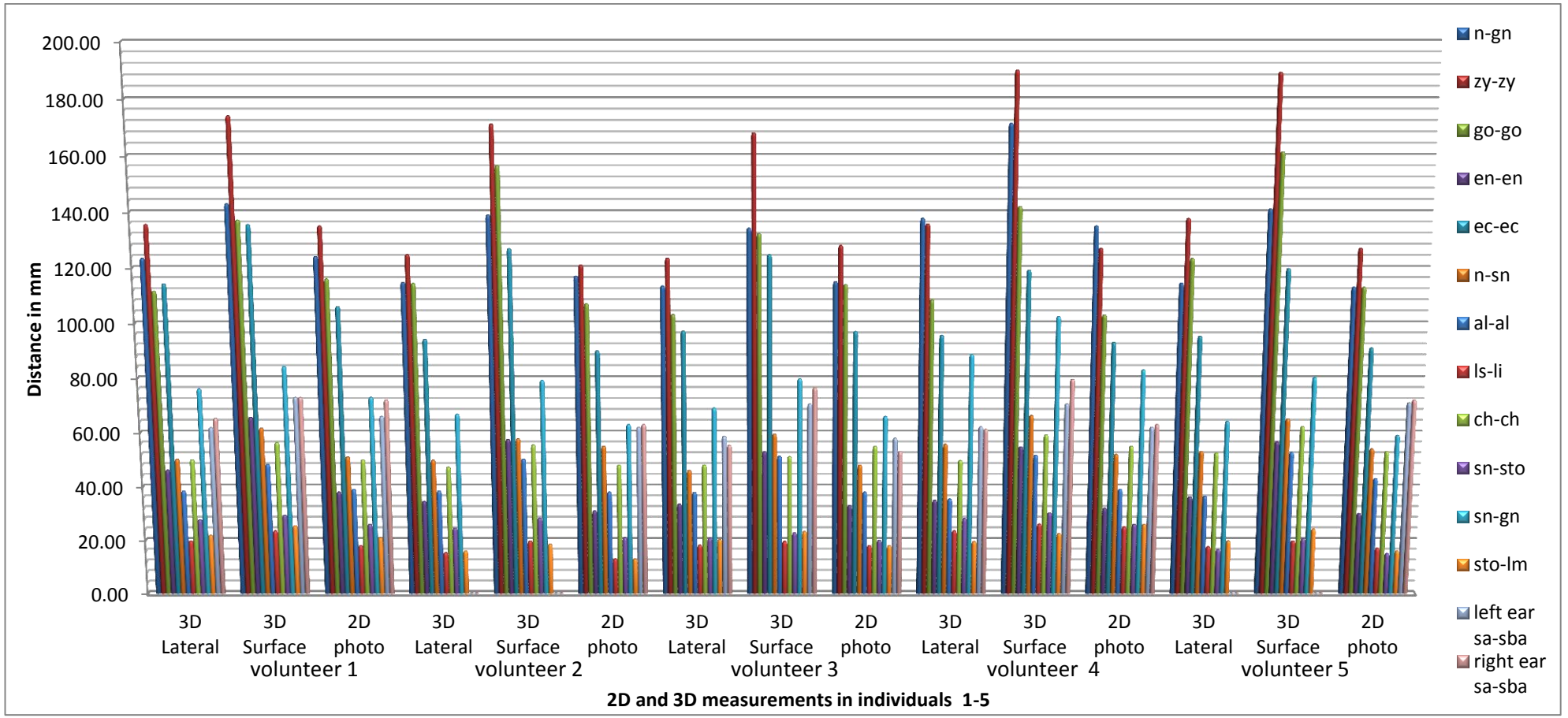


Figure 42. Graphical representation of the comparison between 2D and 3D measurements in individuals 1-5, based on Table 14.

Table 15. Results of the comparison between the craniofacial measurements in 2D and 3D images, including later and surface distance. Values shown are in millimetres.

	Volunteers														
	6			7			8			9			10		
Measurements	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo	3D Lateral	3D Surface	2D photo
n-gn	117.30	139.57	118.00	113.14	128.95	101.00	122.51	145.34	122.00	113.68	127.13	116.00	113.53	135.74	112.00
zy-zy	138.60	175.13	129.00	130.33	175.87	119.00	124.08	170.40	114.00	129.36	159.22	130.00	136.68	180.11	126.00
go-go	119.42	149.25	114.00	113.52	138.95	106.00	97.15	144.69	91.00	108.15	133.49	107.00	110.22	144.26	104.00
en-en	30.59	43.47	28.00	33.49	51.15	28.00	31.39	43.43	27.50	32.24	47.09	28.00	35.37	49.58	30.00
ec-ec	90.29	103.72	86.00	85.31	106.22	80.00	87.49	106.62	82.00	95.48	117.82	95.00	94.15	113.99	90.00
n-sn	46.56	59.11	51.00	44.30	56.62	42.00	56.76	72.11	58.00	51.34	59.69	57.00	50.35	60.57	50.00
al-al	35.35	44.28	37.00	26.42	38.15	29.00	29.54	41.67	34.00	31.57	39.27	27.50	31.89	43.42	33.00
ls-li	20.70	26.27	19.00	16.69	17.67	15.00	23.49	32.40	23.00	23.01	29.02	20.00	20.92	23.61	20.00
ch-ch	49.63	58.17	48.00	44.29	51.84	45.00	44.80	51.11	45.00	50.94	56.38	48.00	50.59	59.33	50.00
sn-sto	24.23	30.67	20.00	19.88	20.93	15.00	21.41	27.10	19.00	20.84	23.96	19.00	24.52	27.10	21.00
sn-gn	76.17	88.25	66.00	66.47	73.89	59.00	65.57	80.72	64.00	62.89	73.35	59.00	67.12	76.39	62.00
sto-lm	19.25	22.08	18.00	18.17	21.59	15.00	21.59	29.41	18.00	19.19	23.10	16.00	17.46	21.59	18.00
left ear sa-sba	66.16	74.37	61.00	53.76	67.09	54.00	63.15	75.07	52.00	58.42	75.40	56.00	54.32	64.92	54.00
right ear sa-sba	66.33	78.21	60.00	54.74	64.36	55.00	60.76	69.61	53.00	57.75	67.33	59.00	60.07	70.49	54.00

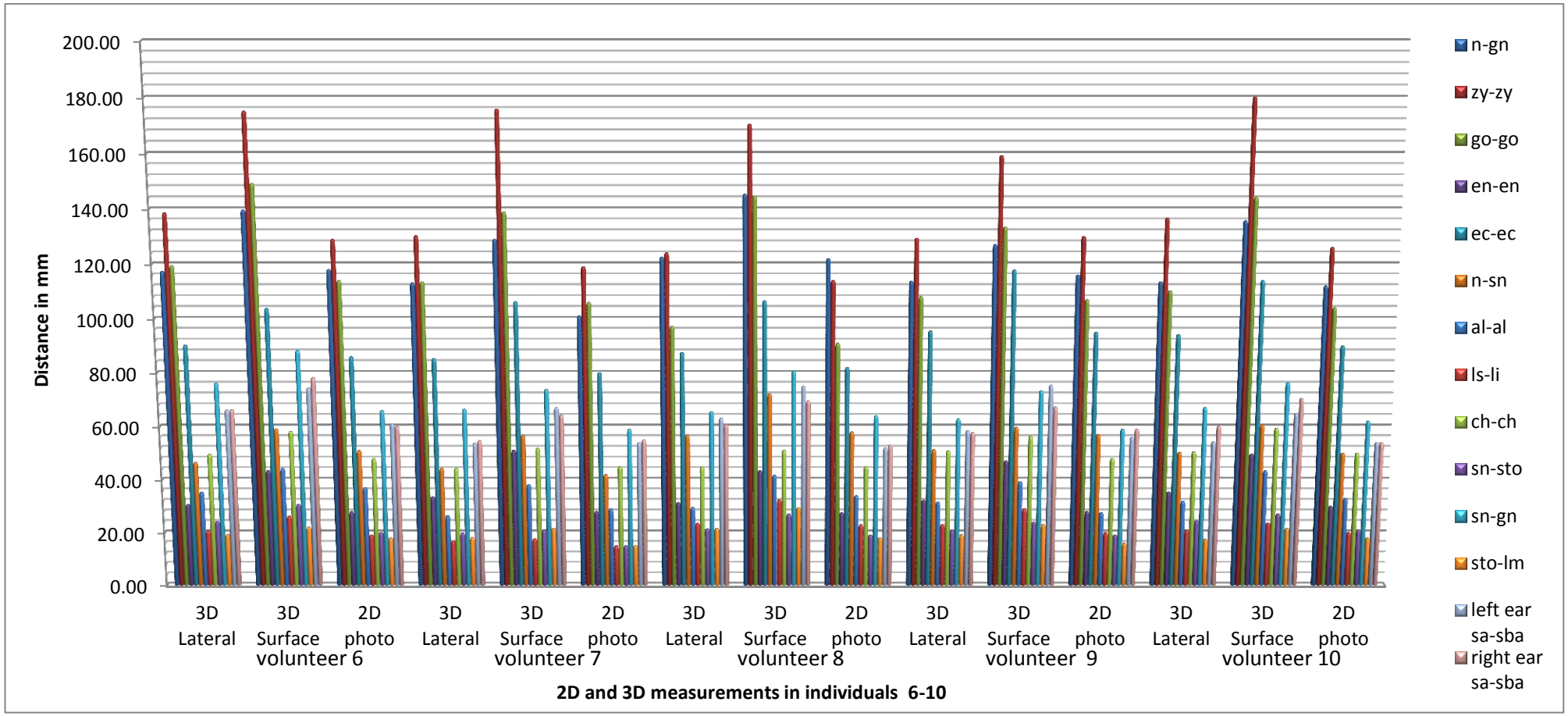


Figure 43. Graphical representation of the comparison between 2D and 3D measurements in individuals 6-10, based on Table 15.

4.4. Reproducibility of craniofacial measurements

4.4.1. Introduction

Reproducibility of the craniofacial measurements can be defined as the ability to obtain the same result, with the same (or different) examiner over a period of time (usually days to months). This concept represents one of the most fundamental principles of the anthropometry and must be investigated thoroughly, prior to conducting a final study.

The accurate location of the soft-tissue facial anthropometrical landmarks and subsequent measurements are not trivial tasks to perform on a living individual. An even higher level of complexity exists when this procedure is performed on a digital 3D image. The landmarks are usually palpated for accurate allocation, which is not possible with digital images. Palpation is especially required for measurement of the landmarks located on or around bony prominences, which are more reproducible, such as left and right zygion (zy) and gonion (go). However, this is not possible with 3D images. The inability to reach accurate location of specific landmarks, may introduce an error in subsequent measurements. Some landmarks may be less reliable because the 3D scanning process does not efficiently capture eyes (pupils), hair and sometimes lip area, due to technical limitations of the laser capturing method. The landmarks that may introduce an error in measurements include the following: trichion (tr), left and right exocanthion (ex) and endocanthion (en), labiale inferius (li), labiale superius (ls), left and right cheilion (ch) and stomion (sto). In contrast, landmarks that were easier to find, included the following: pronasale (prn), left and right alare (al) and nasion (n) all located at the nose area; gnation (gn), pogonion (pg) that are located at the chin area; sublabiale (sl) that is located at the lips area; glabella (g)) that is located at the forehead; left and right endocanthion (en) that are located at the eyes area and all the landmarks located on the ear: left and right superaurale (sa), subaurale (sba), postaurale (pa) and tragion (t).

This present study tested the reproducibility of the craniofacial landmarks allocation on a small subset of individuals by calculating derived distances. The results of this study were used as a proof of concept and provided a basis for collection of a larger dataset.

4.4.2. Materials and Methods

In order to validate the reproducibility of the facial measurements, thirteen 3D images were analysed for a full set of 32 facial landmarks twice, as detailed in the Chapter 2. The period between the analyses varied from one to six months. All facial landmarks were allocated manually, following the same strict methodology. The Euclidean coordinates for 32 landmarks were exported into Microsoft Excel and 86 distances and ratios were calculated automatically using the formulae for linear and angular distances, as detailed in Chapter 2.

The mean difference (MD) was calculated as the discrepancy between the first and the second measurements $\sqrt{(first \ - second)^2}$. The measurement error (ME) was calculated as the standard deviation of the MD divided by square root of 2 (ME=SD(MD/$\sqrt{2}$).

4.4.3. Results and Discussion

The aim of this study was to evaluate the reproducibility and reliability of 86 facial measurements, obtained from 3D facial images. In digital images, the bony structures lying under the soft tissue are neither visible nor available for palpating. As a result, measurements requiring location of bone-related landmarks (such as gonion, zygion and glabella) may be less reproducible on 3D laser-captured images. An a priori assumption was that measurements generated using landmarks located on the lip and eye areas would generate more variation than measurements involving the nose and ear landmarks (specifically the nasion, pronasale, subnasale and tragion), because these areas were captured with relatively low efficiency by the scanner. In general, the data on landmarks in the eye and lips areas were limited as they were captured with low efficiency. The nasal area landmarks and tragion were the easiest to find because of their defined anatomical location. Due to the location of the trichion (the hairline in the middle of the forehead) and given the issues with scan capture of the hair, that landmark was also expected to show more variation than others.

Figure 44 shows an example of the variation between two observations that generated the minimum difference between most of the first and the second measurements. In contrast, Figure 45 shows an image, which generated the maximum difference between these measurements.

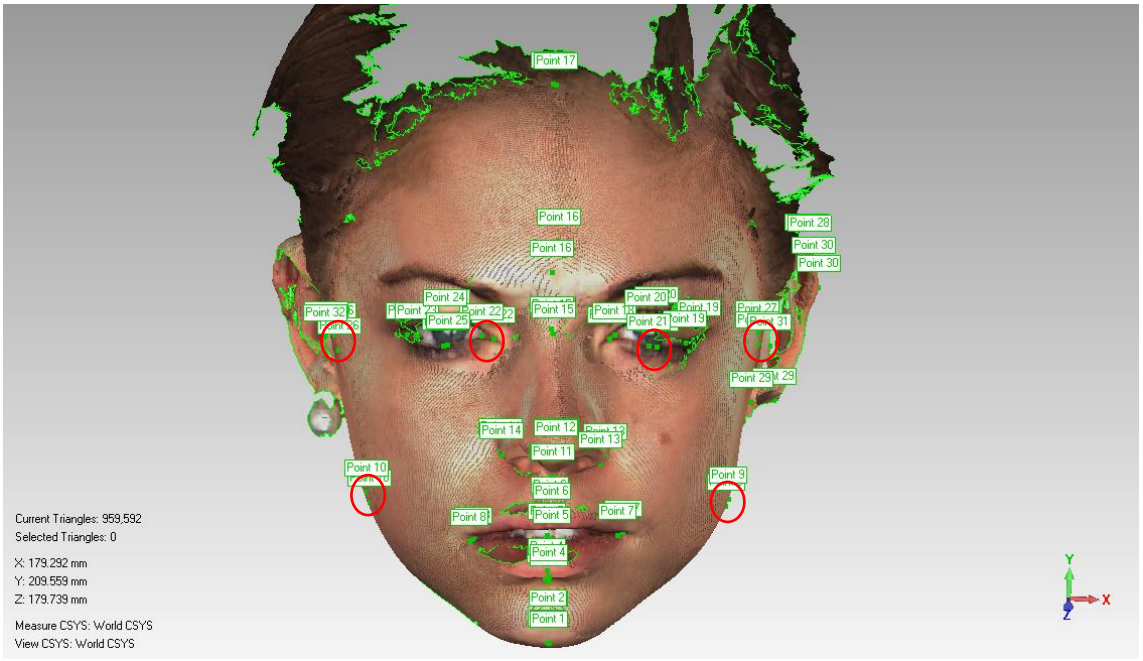


Figure 44. A 3D image, tested twice for location of 32 facial landmarks and generated minimum variance between most of the "old" and the "new" anthropometric measurements. Red circles indicate pairs of landmarks, showing significant difference between two observations.

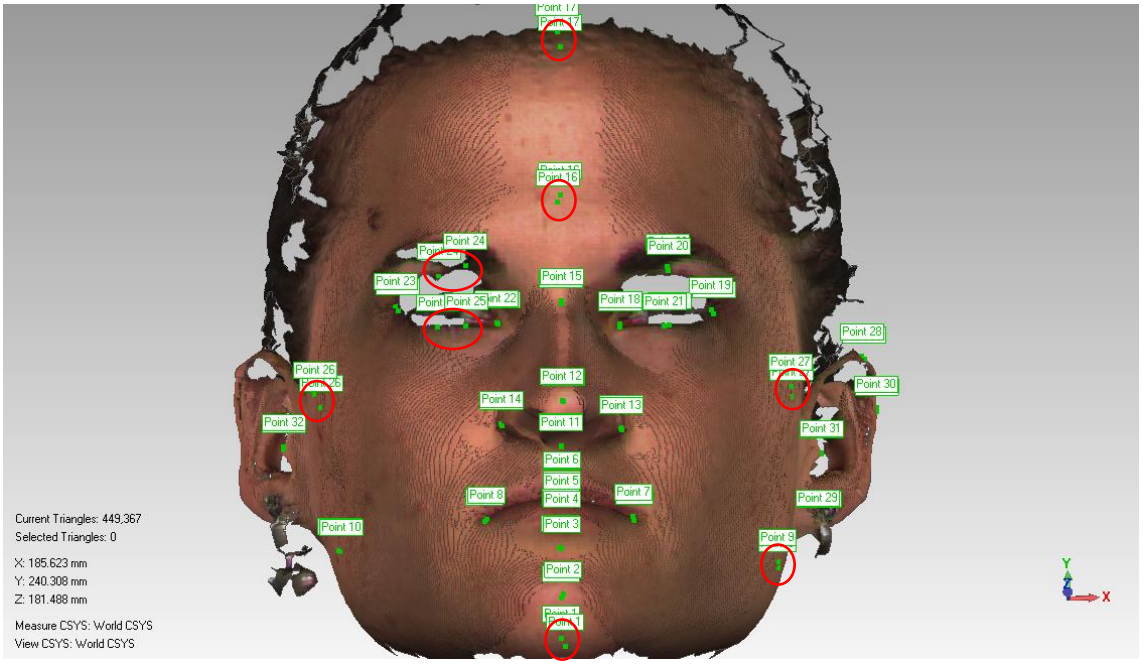


Figure 45. A 3D image, tested twice for location of 32 facial landmarks and generated maximum variance between most of the "old" and the "new" anthropometric measurements. Red circles indicate pairs of landmarks, showing significant difference between two observations. Note the poor coverage of the eye area.

Tables 16, 17 and 18 demonstrate the summary of 54 linear distances, 22 ratios between linear distances and 10 angular distances respectively in 13 facial images. For linear distances, the highest variance was observed between measurements involving paired landmarks, such as gonion and zygion, with two samples generating most of the variability observed. A possible explanation for this variability is poor image quality and general difficulty in finding these landmarks. The analysis of the average values showed that go(r)-zy(r), zy(r)-gn and tr-zy(r) measurements (5.6%) resulted in more than 6 mm MD between the two observations; go(l)-zy(l), and zy(l)-gn (3.7%) resulted in more than 5 mm MD; seven measurements (13%) resulted in more than 4 mm MD; eight measurements resulted in more than 3 mm MD (14.8%); eleven measurements (20.4%) resulted in more than 2 mm MD and 23 measurements (42.6%) resulted in less than 2 mm MD between two observations. The relatively high variation in linear distances involving gonion, zygion and trichion can be explained by the difficulty in accurate location of these landmarks. On the other hand, more than 62% of the measurements resulted in approximately 2 mm (or less) difference between the two observations.

Table 16. A summary of 54 linear measurements for thirteen 3D images with detailed average, minimum and maximum values.

	tr-gn	zy-zy	go-go	prn-gn	t(L)-gn	t(L)-prn	t(L)-n	t(L)-g	t(L)-tr	t(R)-tr	t(R)-gn	t(r)-t(l)	tr-zy(l)
Average	2.46	2.30	2.68	3.40	1.39	0.91	1.07	1.55	1.41	1.37	1.31	0.81	4.62
Minimum	0.10	0.55	0.25	0.01	0.25	0.02	0.04	0.12	0.07	0.40	0.02	0.01	0.07
Maximum	7.10	6.55	9.24	6.36	4.41	3.60	5.34	5.47	6.33	3.21	3.24	3.71	10.69
	tr-zy(r)	n-zy(l)	n-zy(r)	zy(l)-gn	zy(r)-gn	tr-n	g-tr	g-gn	n-gn	n-sto	sto-gn	n-prn	n-sn
Average	6.33	3.48	3.06	5.53	6.25	2.06	4.20	4.86	2.67	1.38	2.71	1.13	1.90
Minimum	0.76	0.37	0.59	1.02	0.40	0.04	0.92	0.03	0.12	0.24	0.20	0.01	0.43
Maximum	14.44	12.32	6.67	14.00	15.13	5.59	10.08	13.60	6.32	3.78	7.42	2.44	3.25
	sn-prn	sn-gn	sl-gn	gn-go(l)	gn-go(r)	prn-go(l)	prn-go(r)	go(l)-tr	go(r)-tr	go(r)-zy(r)	go(l)-zy(l)	zy(l)-al(l)	zy(r)-al(r)	al-al
Average	1.10	3.13	2.98	3.36	4.42	2.80	4.52	3.51	3.86	6.72	5.66	4.34	3.86	0.77
Minimum	0.12	0.40	0.50	0.10	0.05	0.16	0.77	1.10	0.93	0.22	0.54	0.31	1.09	0.11
Maximum	4.42	7.74	7.34	13.84	15.36	6.23	13.35	7.03	12.31	13.27	13.67	14.22	8.62	1.94
	prn-al(l)	prn-al(r)	g-pg	en-en	ex-ex	en(l)-ex(l)	en(r)-ex(r)	ps(l)-pi(l)	ps(r)-pi(r)	sa(l)-sba(l)	t(l)-pa(l)	ch-ch	ls-sto	li-sto
Average	1.21	1.26	4.47	1.59	1.79	1.40	2.51	1.17	1.60	2.37	1.33	2.30	0.76	1.17
Minimum	0.06	0.12	0.31	0.04	0.14	0.06	0.58	0.08	0.07	0.32	0.04	0.18	0.05	0.05
Maximum	3.02	3.37	9.66	5.37	7.40	7.32	6.93	3.13	3.55	6.94	3.17	7.31	2.67	3.32

Table 17. A summary of 22 ratios between linear distances for thirteen 3D images with detailed average, minimum and maximum values.

	Forehead height ratio (tr-n*100/go(r)-go(l))	Upper face height ratio(n-sn*100/go(r)-go(l))	Lower face height ratio (sn-gn*100/go-go)	Mandible index: (sto-gn)*100 /(go-go)
Average	2.16	1.78	2.65	2.13
Mimimum	0.08	0.14	0.19	0.44
Maximum	6.09	4.94	6.29	5.08
	Interendocanthion distance ratio (en-en*100/al-al)	Upper face height ratio (n-sn*100/sn-gn)	Total anterior face height ratio (tr-gn*100/zy-zy)	Mandible – Face width ratio (go-go*100/(zy-zy)
Average	3.86	5.35	3.40	2.85
Mimimum	0.10	0.48	1.17	0.11
Maximum	9.48	16.11	5.49	6.53
	go-go*100/ex-ex	Exocanthal index ex(R)-en(R)*100/en(L)-ex(L)	Intercanthal index en(R)-en(L)*100/ex(R)-ex(L)	Face height index: (n-gn)*100/(tr-gn)
Average	4.65	6.77	2.44	0.88
Mimimum	0.25	0.42	0.28	0.02
Maximum	12.95	16.26	10.09	2.84
	Nose-face height index n-sn/n-gn	Nose-face width index al-al/zy-zy	Nasal tip prostrusion - width index sn-prn/al-al	Nasal Tip Protrusion –Nose height index: (sn-prn)x100/(n-sn)
Average	1.65	0.46	2.98	1.35
Mimimum	0.11	0.00	0.01	0.08
Maximum	4.94	1.25	11.45	6.94
	Anterior face height 1 (n-gn*100/go-go)	Anterior face height 2 ratio (n-gn*100/zy-zy)	Upper face height ratio (n-sn*100/zy-zy)	
Average	2.57	2.97	1.25	
Mimimum	0.19	0.09	0.03	
Maximum	8.73	6.25	3.08	
	Mouth width ratio (ch-ch*100/en-en)	Forehead/lower face height index (Tr-g*100/sn-gn)	Nasal Index al-al/n-sn	
Average	4.46	8.02	2.12	
Mimimum	0.20	2.37	0.14	
Maximum	11.31	17.40	6.75	

Table 18. A summary of 10 angular distances for thirteen 3D images with detailed average, minimum and maximum values.

	Nasal tip angle (n-prn-sn)	Nasal vertical prominence (tr-prn-gn)	Transverse nasal prominence (zy(l)-prn-zy(r))
Average	2.79	1.12	2.41
Mimimum	0.19	0.04	0.09
Maximum	6.84	3.59	6.15
	Nasolabial angle (prn-sn-ls)	**Nasofrontal angle (g-n-prn)**	**Forehead nasal angle (tr-n-prn)**
Average	3.75	2.32	1.14
Mimimum	0.08	0.45	0.17
Maximum	9.28	4.20	4.06
	Chin prominence (go(l)-gn-go(r))	**Transverse nasal prominence 2 (t(l)-prn-t(r))**	**Nasion depth angle (zy(l)-n-zy(r))**
Average	2.25	0.38	2.52
Mimimum	0.24	0.01	0.21
Maximum	6.89	1.91	7.43
	Nasomental angle (n-prn-pg)		
Average	0.80		
Mimimum	0.13		
Maximum	1.63		

The graphical representation of the results is shown in Figure 46. The highest measurement error (ME) was observed in measurements involving paired landmarks (gonion and zygion), while the lowest was in measurements involving the nasal area landmarks and tragion (ear), as summarised in Figure 47. The analyses of the results involved only basic descriptive statistics, as a more comprehensive analysis was beyond the scope of this project.

The comparison between the craniofacial ratios revealed that forehead/lower face height index (Tr-g*100/sn-gn) demonstrated the highest variance (8.02 mm), which can be accounted for three (3) samples that demonstrated significantly higher values than the rest of the dataset. This is most likely due to variation in location of the trichion, which can be covered by hair. The exocanthal index also showed relatively high variance (6.77 mm) due to poor capture of the eye area, confirming the original assumption of poor laser capture. However, the majority of the ratios (68.2%) showed MD of less than 3 mm and ME of less than 1.5 mm (Figures 46 and 47).

Evaluation of the reproducibility of the angular distances revealed that the nasolabial angle (prn-sn-ls) showed the highest variability of 3.75 degrees among the measurements. One sample showed significantly higher variance due to poor digital capture of the lip area, which affected the accurate location of the labiale superius (ls). The rest of the angular measurements showed MD of 0.38 to 2.79 degrees. The ME range was also low, with values between 0.39 mm to 1.56 mm.

The overall low variance in angular distances can be explained by the nature of these measurements. In 3D space, the angular distance is mostly affected by the z-axis (the depth), which is usually unaffected during the landmark location process. However, the variance in landmarks location at x and y-axes can also affect the angular distance. Table 19 shows an example of artificial manipulation with x, y and z coordinates of the prn landmark. The initial coordinate was either appended or subtracted by 5 mm or 10 mm and MD from the original value was calculated. Notably, the subtraction in both y and z coordinates had a more pronounced effect than addition, while for the x coordinate it was the opposite. This phenomenon can potentially be explained by the facial anatomy (e.g. skin folds).

Figure 46. Three plots showing distribution of the mean difference (MD) values as an average of thirteen samples for linear distances (left plot), ratios between linear distances (middle plot) and angular distances (right plot).

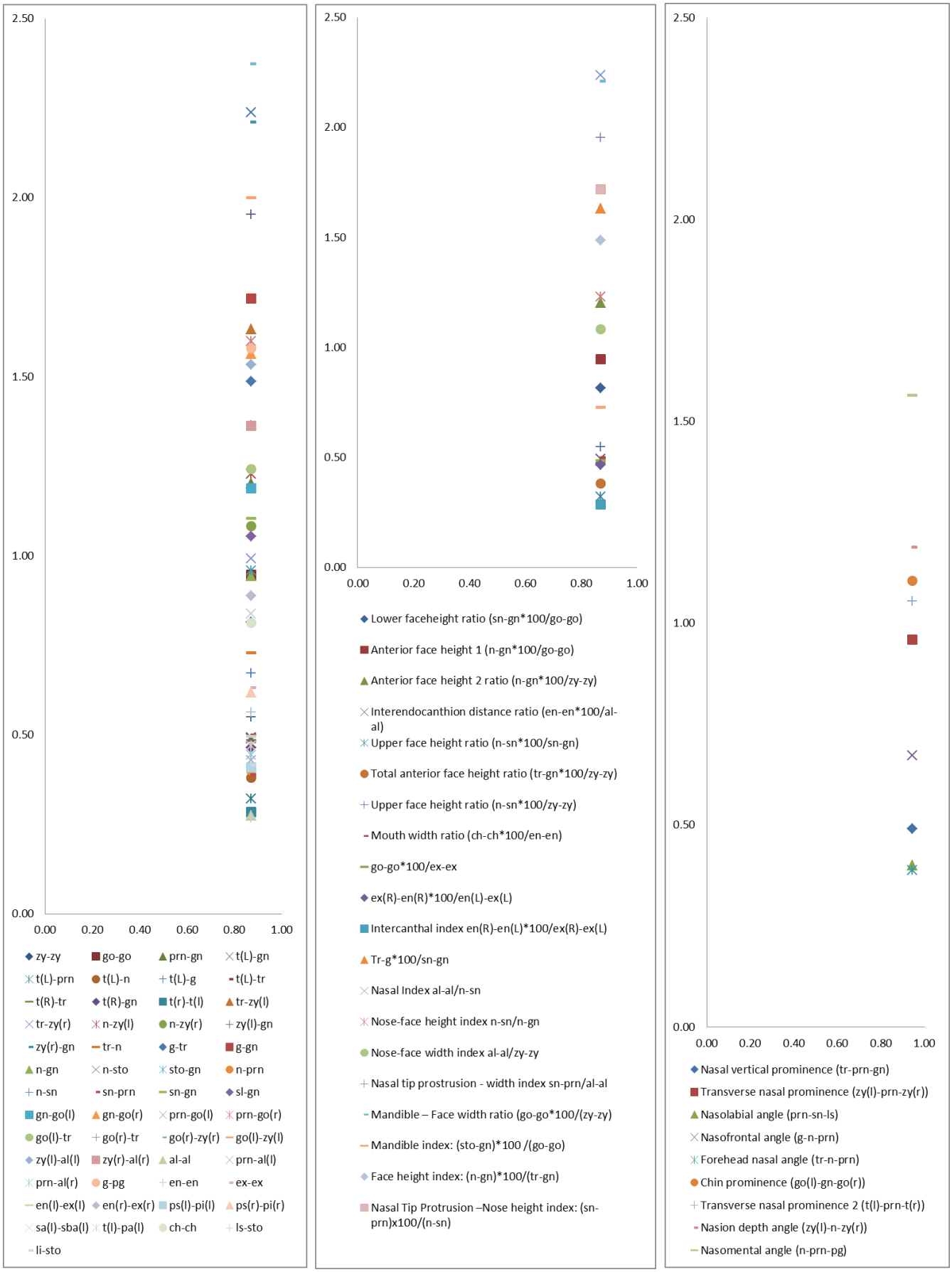


Figure 47. Three plots showing distribution of the measurement error (ME) values as an average of thirteen samples for linear distances (left plot), ratios between linear distances (middle plot) and angular distances (right plot).

Table 19. An artificial manipulation with original coordinates of the 'prn' landmark, showing mean difference (MD) comparing to the initial angular distance. The values are shown in degrees. The landmarks that were mostly affected by manipulation with coordinates location are highlighted in yellow.

		(n-prn-sn)	tr-prn-gn)	(zy(l)-prn-zy(r))	(prn-sn-ls)	(g-n-prn)	(tr-n-prn)	(t(l)-prn-t(r))	(n-prn-pg)
	Initial angular distance	**101.03**	**139.66**	**80.87**	**124.78**	**151.08**	**162.12**	**65.38**	**128.33**
MD from initial distance	prn z axis plus 5mm	16.31	6.21	3.30	10.20	6.17	5.95	2.31	10.36
	prn z axis minus 5mm	28.80	12.15	6.43	17.09	11.80	11.49	4.48	19.90
	prn x axis plus 5mm	11.89	6.05	3.28	5.08	5.62	5.79	2.26	9.63
	prn x axis minus 5mm	2.66	1.21	0.56	1.38	1.31	2.65	0.32	2.32
	prn y axis plus 5mm	7.10	0.07	0.92	12.45	3.86	4.94	0.00	2.61
	prn y axis minus 5mm	16.96	2.08	2.50	22.60	5.70	5.69	0.36	1.64
MD from initial distance	prn z axis plus 10mm	38.00	12.64	6.79	25.44	12.78	11.70	4.76	21.39
	prn z axis minus 10mm	60.36	24.26	12.86	37.18	23.45	22.24	8.98	39.60
	prn x axis plus 10mm	17.62	10.82	6.14	6.47	9.34	8.81	4.26	16.24
	prn x axis minus 10mm	4.48	2.30	1.11	2.65	2.32	4.39	0.63	4.21
	prn y axis plus 10mm	15.02	0.46	1.74	26.22	8.31	10.76	0.03	6.15
	prn y axis minus 10mm	34.10	4.21	4.94	45.72	11.75	11.72	0.72	3.52

4.4.4. Conclusion

It is important to demonstrate reproducibility of facial measurements taken by different analysts or the same examiner at different times. While this topic has been extensively investigated in the traditional direct anthropometry, the number of publications exploring this issue in 3D imaging systems is relatively limited [360-362]. There are two current studies of the reproducibility of facial soft-tissue landmarks with 3D laser scans [41, 363]. In addition, all the current articles deal with the reproducibility of landmark locations, rather than actual facial measurements (specifically the inter or intra-observer variance in location of the x, y and z coordinates). The results of an inter-observer study with a 3D laser scanning system showed MD values between 0.39 mm to 1.49 mm for each landmark, while approximately 50% of the landmarks were reproducible to 1 mm or less and approximately 11% to more than 1 mm [41]. Another study resulted in variance of 0.19 mm to 3.49 mm with a ME range of 0.55 mm to 3.34 mm [41, 363]. These results are somewhat surprising, as the direct anthropometrical measurements error between two observations has been shown to be up to 2 mm in morphologically normal subjects, while being considered the most accurate method to

date [17]. In addition, most of the craniofacial landmarks are not defined by a fixed precise location on the face and can vary slightly between measurements, even when performed by an experienced examiner. Location of the craniofacial landmarks on 3D images is significantly more complicated than direct measurement and is expected to produce a higher variance between two independent observations. It should also be emphasized that linear distances involve location of two landmarks, while angular distances require three landmarks. This means that the demonstrated upper variance limit of 3.34 mm should be increased two or three fold (6.68-10.02 mm) for actual measurements previously demonstrated. In summary, it can be concluded that the mean distance difference in the present study successfully produced lower variance and, as a result, higher reproducibility, compared to previous published studies.

4.5. Assessment of normal distribution of the craniofacial measurements, including sex and ethnic differences.

4.5.1. Introduction

The variation in craniofacial morphology between genders and ethnic groups has been systematically explored since the end of 19th century [19]. Despite the visual differences in the craniofacial morphology among races, most research has focused on the Caucasian population and particularly on European populations [17, 35]. In the last few years several studies of anthropometric measurements in other populations have been published. The most comprehensive study compared a set of 14 facial distances among a group of 30 males and females from each of 25 countries [28]. Interest in finding the differences in craniofacial morphology between various ethnicities has practical application in plastic surgery, specifically in craniomaxillofacial and rhinoplastical intervention for restoring dysmorphology or improving aesthetics, where the knowledge of normal proportions of the head and face is important. Due to the limited information currently available, more extensive research involving a greater number of individuals from various ethnicities is required to establish a reliable database of normal variation in craniofacial morphology.

Any conclusions regarding the significance of variance between population groups rely primarily on statistical analyses of the data. Most of the parametric statistical tests, such as correlation, regression and analysis of variance are based on the assumption that the data follow a Gaussian (normal) distribution [364, 365]. As a result, prior to performing statistical analyses, the dataset should be checked for normality. Numerous publications however, argue that in a real dataset there is no such thing as true normality [366, 367]. In addition, with a large sample size (hundreds of observations) non-normal distribution can be ignored [368, 369]. Nevertheless, the test for normal distribution of data is essential to find any extreme values (potential errors), helping to make the outcome of statistical tests more accurate and reliable.

The main aim of this study was to test the craniofacial distances for outliers or extreme values from the normal distribution (for each sex), which may represent potential errors in measurements. Following the normality test, several measurements were analysed in order to find craniofacial regions that contribute most of the differences between the five main population groups, according to sex. This evaluation represents an important

quality control step in data analysis and adds valuable information of the normal craniofacial measurement range to the published data.

The analysis of distribution of all 92 measurements in several population groups was considered unnecessary and beyond the scope of this project. As a result, the data were analysed for a few measurements that contribute most significantly to the overall variance. These values were represented by principal components (PC).

Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set. This reduction is accomplished by identifying directions, visualised as eigenvectors, along which the variation in the data is expected to be maximal. By using a few components, each sample, genotype or trait can be represented by significantly less numbers instead of possibly thousands of variables. Samples, genotypes or traits can then be plotted to assess similarities and differences between them and determine whether they can be grouped. Several principal components obtained in this study were subsequently used for analysing their association with genetic markers, thus significantly reducing the time for statistical analysis. Furthermore, performing PCA on the craniofacial measurements allowed to produce "morphological vectors" represented by similar anthropometrical distances, which represent the facial appearance in a comprehensive fashion (as discussed in Section 4.6). As a result, the association analysis of craniofacial principal components may reveal more significantly associated genetic markers, hence better prediction of the facial appearance.

4.5.2. Materials and Methods

All the statistical analyses were performed using PASW Statistics (SPSS version 19, IBM Corp.). The normal distribution of the data was checked by analysing skewness and kurtosis of the dataset, as well as by Shapiro-Wilk (S-W) test. The tests were performed for each measurement, sex separated. The distribution was considered normal if z-scores of skewness and kurtosis were between -2.58 and 2.58 and p-value for S-W test greater than 0.01. The Kolmogorov-Smirnov test was not relevant for this study, as it is less powerful and more sensitive to extreme values than the S-W test. The outliers were defined as either ≥3x interquartile range (IQR) above the third quartile or ≥3×IQR below the first quartile.

Principal component analyses were performed on all the linear and angular measurements, including the ratios between these measurements, using PASW Statistics (version 19, IBM Corp.). The plots were prepared in PASW Statistics and edited in Microsoft Excel.

4.5.3. Results and Discussion

Performing the test for normality on the measurements allowed testing a hypothesis of normal distribution of the data, which is an initial step prior to statistical analysis. It also helped to identify outliers, represented by extreme values. These values were caused either by wrong location of facial landmarks, which led to errors in calculations and were identified and corrected in the final dataset.

The z-scores for skewness and kurtosis values for symmetrical Gaussian distribution, were expected to lie between -1.96 and 1.96 at p-value > 0.05. However, in a large dataset (≥200 observations) this range can be changed to ±2.58 at p-value >0.01 [368, 369]. Given the current sample size, these values were applied for analysis of the dataset.

The normality test performed on direct cranial measurements and respective ratios in the whole dataset (all ethnicities together), revealed that two direct measurement (G-Op and Eu-Eu) showed normal distribution in both genders, according to the S-W test p-values (Tables 20 and 21). The v-gn/height ratio was distributed normally in females only. The Eu-Eu distance in females displayed positive skewness and less significant p-value than G-Op. The V-Gn distance was negatively skewed, while the eu-eu/v-gn ratio was positively skewed in both sexes, failing the normality test. In all the tests of direct measurements, no significant outliers from normal distribution were detected.

Table 20. Normality tests for direct craniofacial measurements and ratios in females. a. Lilliefors Significance Correction, *. This is a lower bound of the true significance. Values, within the limits of normal distribution highlighted in red.

Tests of Normality - females

	Kolmogorov-Smirnov[a]			Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Statistic	df	Sig.	Skewness	Kurtosis
V-Gn (Craniofacial height) mm	.055	314	.021	.979	314	.000	-3.746	4.283
Eu-Eu (Head Width) mm	.047	314	.087	.990	314	.039	2.590	0.375
G-Op (Head Length) mm	.034	314	.200*	.997	314	.805	-0.258	-0.625
Cephalic Index	.063	314	.004	.983	314	.001	3.642	0.838
Head width – Craniofacial height ratio (eu-eu*100/v-gn)	.055	314	.022	.973	314	.000	4.840	5.925
v-gn*100/height	.032	314	.200*	.995	314	.449	0.656	1.477

Table 21. Normality tests for direct craniofacial measurements and ratios in males. a. Lilliefors Significance Correction, *. This is a lower bound of the true significance. Values, within the limits of normal distribution highlighted in red.

Tests of Normality - males

	Kolmogorov-Smirnov[a]			Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Statistic	df	Sig.	Skewness	Kurtosis
V-Gn (Craniofacial height) mm	.053	209	.200*	.979	209	.004	-3.144	1.620
Eu-Eu (Head Width) mm	.045	209	.200*	.993	209	.401	0.554	1.679
G-Op (Head Length) mm	.066	209	.026	.989	209	.101	-1.511	-1.178
Cephalic Index	.095	209	.000	.972	209	.000	2.901	-0.453
Head width – Craniofacial height ratio (eu-eu*100/v-gn)	.057	209	.098	.979	209	.003	2.823	3.703
v-gn*100/height	.051	209	.200*	.982	209	.009	1.089	3.718

Normality tests for linear measurements revealed that majority of the facial distances obtained from 3D images were distributed normally after removing several extreme outliers. Specifically, three distances displayed deviations from normal distribution in females: prn-al(l) prn-al(r) and ps(r)-pi(r) as did one in males: tr-gn (Tables 22 and 23). This may be explained by either poor laser capture of the eye and nose areas and difficulty in finding an accurate location of the tragion (hindered by hairline) or by existence of true outliers, representing actual data points.

Table 22. Normality tests for linear facial measurements in females without ethnic separation. a. Lilliefors Significance Correction, *This is a lower bound of the true significance. Values, within the limits of normal distribution highlighted in red.

Tests of Normality - female

	Kolmogorov-Smirnov[a]			Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Statistic	df	Sig.	Skewness	Kurtosis
tr-gn	.036	241	.200*	.994	241	.511	-0.522	-0.760
zy-zy	.056	241	.063	.994	241	.400	0.446	-0.513
go-go	.044	241	.200*	.995	241	.686	1.057	-0.747
prn-gn	.028	241	.200*	.994	241	.509	0.433	-1.010
t(l)-gn	.041	241	.200*	.995	241	.665	0.561	-0.426
t(l)-prn	.049	241	.200*	.987	241	.027	-0.815	-2.157
t(l)-n	.056	241	.061	.991	241	.134	-0.108	-1.978
t(l)-g	.027	241	.200*	.997	241	.912	0.363	-0.862
t(l)-tr	.031	241	.200*	.991	241	.159	-1.191	-1.218
t(r)-tr	.035	241	.200*	.996	241	.747	0.548	-0.737
t(r)-gn	.026	241	.200*	.998	241	.997	-0.115	-0.067
t(r)-t(l)	.033	241	.200*	.994	241	.450	1.146	0.833
tr-zy(l)	.042	241	.200*	.994	241	.510	-0.599	1.298
tr-zy(r)	.041	241	.200*	.993	241	.325	-0.573	1.304
n-zy(l)	.027	241	.200*	.998	241	.980	0.306	0.029
n-zy(r)	.030	241	.200*	.995	241	.681	0.860	-0.551
zy(l)-gn	.031	241	.200*	.996	241	.790	-0.357	-0.093
zy(r)-gn	.042	241	.200*	.995	241	.557	-0.618	-1.054
tr-n	.039	241	.200*	.993	241	.370	-0.828	0.420
g-tr	.048	241	.200*	.994	241	.459	-0.070	-1.221
g-gn	.040	241	.200*	.993	241	.360	-1.318	0.192
n-gn	.044	241	.200*	.995	241	.590	0.299	-0.958
n-sto	.038	241	.200*	.997	241	.894	0.452	-0.490
sto-gn	.042	241	.200*	.996	241	.751	1.013	-0.353
n-prn	.038	241	.200*	.995	241	.639	1.096	0.692
n-sn	.049	241	.200*	.991	241	.119	2.019	2.160

sn-prn	.029	241	.200*	.991	241	.167	1.115	1.657
sn-gn	.030	241	.200*	.997	241	.923	-0.115	-0.904
sl-gn	.040	241	.200*	.993	241	.364	1.344	1.234
gn-go(l)	.033	241	.200*	.992	241	.185	-1.089	0.446
gn-go(r)	.034	241	.200*	.995	241	.687	-1.459	0.321
prn-go(l)	.058	241	.050	.993	241	.307	0.758	-1.138
prn-go(r)	.045	241	.200*	.994	241	.414	0.287	-0.667
go(l)-tr	.048	241	.200*	.994	241	.415	-0.796	-0.647
go(r)-tr	.033	241	.200*	.998	241	.981	0.000	-0.263
go(r)-zy(r)	.048	241	.200*	.993	241	.277	0.892	-0.103
go(l)-zy(l)	.046	241	.200*	.994	241	.465	-0.796	-1.016
zy(l)-al(l)	.042	241	.200*	.987	241	.026	1.083	2.782
zy(r)-al(r)	.031	241	.200*	.990	241	.112	0.924	3.304
al-al	.048	241	.200*	.990	241	.082	2.510	0.740
prn-al(l)	.047	241	.200*	.981	241	.003	2.892	0.234
prn-al(r)	.061	241	.028	.983	241	.005	2.643	-0.615
g-pg	.046	241	.200*	.990	241	.099	-1.408	-0.131
en-en	.052	241	.200*	.993	241	.320	0.847	-0.827
ex-ex	.034	241	.200*	.994	241	.482	1.013	-1.176
en(l)-ex(l)	.051	241	.200*	.995	241	.664	-0.490	-0.756
en(r)-ex(r)	.040	241	.200*	.993	241	.273	1.490	1.058
ps(l)-pi(l)	.036	241	.200*	.992	241	.180	2.127	1.913
ps(r)-pi(r)	.050	241	.200*	.984	241	.009	3.121	2.628
sa(l)-sba(l)	.048	241	.200*	.993	241	.308	1.650	0.987
t(l)-pa(l)	.051	241	.200*	.993	241	.297	0.268	1.340
ch-ch	.042	241	.200*	.995	241	.610	-0.357	-1.026
ls-sto	.047	241	.200*	.989	241	.064	1.752	-0.587
li-sto	.038	241	.200*	.996	241	.724	0.567	0.872

Table 23. Normality tests for linear facial measurements in males without ethnic separation. a. Lilliefors Significance Correction, *This is a lower bound of the true significance. Values, within the limits of normal distribution highlighted in red.

Tests of Normality - male								
	Kolmogorov-Smirnov[a]			Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Statistic	df	Sig.	Skewness	Kurtosis
tr-gn	.094	129	.008	.969	129	.005	-1.197	0.502
zy-zy	.061	129	.200*	.989	129	.425	-2.100	-1.177
go-go	.042	129	.200*	.994	129	.851	0.182	-0.993
prn-gn	.057	129	.200*	.992	129	.671	-0.244	0.063
t(l)-gn	.053	129	.200*	.983	129	.100	-1.770	2.013
t(l)-prn	.062	129	.200*	.978	129	.031	-1.856	2.888
t(l)-n	.061	129	.200*	.978	129	.036	-2.337	3.236

t(l)-g	.065	129	.200*	.991	129	.582	-0.544	0.079
t(l)-tr	.044	129	.200*	.992	129	.684	0.673	-0.808
t(r)-tr	.038	129	.200*	.993	129	.820	-0.633	-0.951
t(r)-gn	.051	129	.200*	.992	129	.691	-0.611	0.223
t(r)-t(l)	.041	129	.200*	.991	129	.587	-1.001	-0.282
tr-zy(l)	.071	129	.186	.986	129	.200	0.483	0.486
tr-zy(r)	.066	129	.200*	.988	129	.347	0.788	1.073
n-zy(l)	.049	129	.200*	.993	129	.762	0.409	-0.405
n-zy(r)	.054	129	.200*	.988	129	.317	1.182	-1.020
zy(l)-gn	.034	129	.200*	.994	129	.857	0.332	0.362
zy(r)-gn	.052	129	.200*	.983	129	.114	-0.982	-0.005
tr-n	.043	129	.200*	.991	129	.575	0.302	-1.131
g-tr	.062	129	.200*	.977	129	.030	1.381	-0.870
g-gn	.060	129	.200*	.984	129	.124	-1.018	-1.550
n-gn	.056	129	.200*	.991	129	.567	0.198	-0.153
n-sto	.057	129	.200*	.989	129	.426	0.526	-0.484
sto-gn	.052	129	.200*	.988	129	.332	1.767	1.019
n-prn	.054	129	.200*	.987	129	.241	1.565	0.270
n-sn	.059	129	.200*	.991	129	.541	0.719	-0.729
sn-prn	.039	129	.200*	.992	129	.664	-0.418	0.207
sn-gn	.033	129	.200*	.995	129	.921	0.875	-0.117
sl-gn	.049	129	.200*	.985	129	.184	1.855	1.471
gn-go(l)	.044	129	.200*	.996	129	.967	-0.363	0.764
gn-go(r)	.073	129	.088	.974	129	.014	-2.352	2.321
prn-go(l)	.055	129	.200*	.990	129	.502	-1.094	-0.048
prn-go(r)	.061	129	.200*	.985	129	.169	0.236	2.047
go(l)-tr	.048	129	.200*	.988	129	.311	0.227	-1.215
go(r)-tr	.049	129	.200*	.993	129	.814	-0.206	0.186
go(r)-zy(r)	.063	129	.200*	.986	129	.229	-0.558	-0.628
go(l)-zy(l)	.101	129	.003	.977	129	.030	-0.895	-0.265
zy(l)-al(l)	.055	129	.200*	.990	129	.477	0.500	-0.987
zy(r)-al(r)	.054	129	.200*	.988	129	.333	0.202	-1.087
al-al	.094	129	.007	.976	129	.021	2.205	-0.509
prn-al(l)	.040	129	.200*	.994	129	.899	-0.019	0.133
prn-al(r)	.062	129	.200*	.989	129	.388	0.891	-0.987
g-pg	.067	129	.200*	.987	129	.236	-0.862	-1.219
en-en	.063	129	.200*	.988	129	.343	0.848	1.068
ex-ex	.055	129	.200*	.991	129	.564	-0.803	-0.216
en(l)-ex(l)	.049	129	.200*	.992	129	.678	0.411	-0.717
en(r)-ex(r)	.065	129	.200*	.992	129	.635	0.852	-0.540
ps(l)-pi(l)	.073	129	.085	.981	129	.067	2.146	-0.234
ps(r)-pi(r)	.046	129	.200*	.993	129	.776	0.709	-0.301
sa(l)-sba(l)	.040	129	.200*	.992	129	.652	0.215	-0.681
t(l)-pa(l)	.050	129	.200*	.987	129	.289	0.410	-0.699
ch-ch	.044	129	.200*	.993	129	.749	-0.956	0.649
ls-sto	.057	129	.200*	.984	129	.126	1.670	-0.393
li-sto	.080	129	.042	.989	129	.427	-0.351	-0.388

Most ratios between linear facial distances showed normal distribution except for upper face height ratio, intercanthal index and nose-face width index in females and "tr-g*100/sn-gn" ratio in males (Tables 24 and 25). These ratios were calculated using several landmarks, which showed more variability in the reproducibility study, due to poor image quality and complexity in allocating these landmarks, as discussed in Section 4.4.

Table 24. Normality tests for facial ratios in females, without ethnic separation. Values, within the limits of normal distribution highlighted in red.

Tests of Normality for ratios - females

	Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Skewness	Kurtosis
Forehead height ratio (tr-n*100/go(r)-go(l))	.995	283	.474	1.302	-0.533
Upper face height ratio (n-sn*100/go(r)-go(l))	.993	283	.188	2.016	1.219
Lower faceheight ratio (sn-gn*100/go-go)	.993	283	.244	1.408	-0.245
Anterior face height 1 (n-gn*100/go-go)	.992	283	.125	1.988	-0.418
Anterior face height 2 ratio (n-gn*100/zy-zy)	.988	283	.020	2.930	0.745
Interendocanthion distance ratio (en-en*100/al-al)	.994	283	.375	1.389	-0.513
Upper face height ratio (n-sn*100/sn-gn)	.995	283	.516	0.722	0.725
Total anterior face height ratio (tr-gn*100/zy-zy)	.991	283	.068	1.396	-0.905
Upper face height ratio (n-sn*100/zy-zy)	.980	283	.001	3.607	3.913
Mouth width ratio (ch-ch*100/en-en)	.994	283	.280	1.144	1.203
go-go*100/ex-ex	.995	283	.511	0.822	-0.407
ex(R)-en(R)*100/en(L)-ex(L)	.989	283	.030	2.295	0.252
Intercanthal index en(R)-en(L)*100/ex(R)-ex(L)	.996	283	.684	0.367	-0.714
tr-g*100/sn-gn	.990	283	.056	2.116	0.333
Nasal Index al-al/n-sn	.984	283	.003	2.987	2.636
Nose-face height index n-sn/n-gn	.994	283	.393	-0.841	1.007
Nose-face width index al-al/zy-zy	.985	283	.004	2.411	4.493
Nasal tip prostrusion - width index sn-prn/al-al	.998	283	.964	0.411	0.645
Mandible – Face width ratio (go-go*100/(zy-zy)	.993	283	.207	-1.608	1.172
Mandible index: (sto-gn)*100 /(go-go)	.992	283	.128	2.137	0.342
Face height index: (n-gn)*100/(tr-gn)	.995	283	.408	0.933	0.368
Nasal Tip Protrusion –Nose height index: (sn-prn)x100/(n-sn)	.992	283	.147	1.894	1.614

Table 25. Normality tests for facial ratios in males, without ethnic separation. Values, within the limits of normal distribution are highlighted in red.

Tests of Normality for ratios- males

	Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Skewness	Kurtosis
Forehead height ratio (tr-n*100/go(r)-go(l))	0.987	174	0.113	0.831	-0.177
Upper face height ratio (n-sn*100/go(r)-go(l))	0.993	174	0.593	-0.086	1.399
Lower faceheight ratio (sn-gn*100/go-go)	0.991	174	0.356	1.379	0.412
Anterior face height 1 (n-gn*100/go-go)	0.991	174	0.327	0.62	0.272
Anterior face height 2 ratio (n-gn*100/zy-zy)	0.995	174	0.873	0.778	-0.313
Interendocanthion distance ratio (en-en*100/al-al)	0.985	174	0.061	2.397	1.522
Upper face height ratio (n-sn*100/sn-gn)	0.996	174	0.886	0.394	0.768
Total anterior face height ratio (tr-gn*100/zy-zy)	0.987	174	0.094	-0.622	-0.318
Upper face height ratio (n-sn*100/zy-zy)	0.989	174	0.179	-0.913	1.468
Mouth width ratio (ch-ch*100/en-en)	0.989	174	0.216	1.203	-0.432
go-go*100/ex-ex	0.991	174	0.302	0.747	1.116
ex(R)-en(R)*100/en(L)-ex(L)	0.994	174	0.759	0.455	-0.387
Intercanthal index en(R)-en(L)*100/ex(R)-ex(L)	0.997	174	0.978	0.096	0.813
tr-g*100/sn-gn	0.969	174	0.001	2.757	-0.751
Nasal Index al-al/n-sn	0.982	174	0.026	2.814	1.632
Nose-face height index n-sn/n-gn	0.993	174	0.529	-1.483	1.049
Nose-face width index al-al/zy-zy	0.983	174	0.03	2.654	1.698
Nasal tip prostrusion - width index sn-prn/al-al	0.995	174	0.844	0.399	1.001
Mandible – Face width ratio (go-go*100/(zy-zy)	0.986	174	0.075	-0.606	-0.501
Mandible index: (sto-gn)*100 /(go-go)	0.979	174	0.011	2.93	1.503
Face height index: (n-gn)*100/(tr-gn)	0.994	174	0.73	0.283	-1.111
Nasal Tip Protrusion –Nose height index: (sn-prn)x100/(n-sn)	0.992	174	0.4	1.362	0.02

The normality test for angular distances revealed that all the measurements distributed normally in both genders, after removing six outliers (Tables 26 and 27). The outliers were caused by wrong location of the landmarks (as discussed in paragraph below) and were subsequently corrected. These results were expected, because angular distances demonstrated better reproducibility between measurements, as discussed in Section 4.4.

Table 26. Normality tests for angular distances in females, without ethnic separation. Values, within the limits of normal distribution are highlighted in red.

Tests of Normality for angular distances - female

	Kolmogorov-Smirnov			Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Statistic	df	Sig.	Skewness	Kurtosis
Nasal tip angle (n-prn-sn)	0.046	276	.200*	0.993	276	0.189	1.645	0.67
Nasal vertical prominence (tr-prn-gn)	0.044	276	.200*	0.992	276	0.165	1.075	0.076
Transverse nasal prominence (zy(l)-prn-zy(r))	0.056	276	0.036	0.992	276	0.144	2.003	-0.112
Nasolabial angle (prn-sn-ls)	0.036	276	.200*	0.995	276	0.565	-1.428	0.741
Nasofrontal angle (g-n-prn)	0.032	276	.200*	0.992	276	0.139	1.834	0.059
Forehead nasal angle (tr-n-prn)	0.051	276	0.076	0.994	276	0.324	0.884	0.799
Chin prominence (go(l)-gn-go(r))	0.042	276	.200*	0.987	276	0.011	2.052	-1.211
Transverse nasal prominence 2 (t(l)-prn-t(r))	0.048	276	.200*	0.994	276	0.374	1.369	0.255
Nasion depth angle (zy(l)-n-zy(r))	0.044	276	.200*	0.987	276	0.016	2.172	1.46
Nasomental angle (n-prn-pg)	0.032	276	.200*	0.995	276	0.605	-0.006	-0.616

Table 27. Normality tests for angular distances in males, without ethnic separation. Values, within the limits of normal distribution are highlighted in red.

Tests of Normality for angular distances - male

	Kolmogorov-Smirnov			Shapiro-Wilk			z-score	
	Statistic	df	Sig.	Statistic	df	Sig.	Skewness	Kurtosis
Nasal tip angle (n-prn-sn)	0.042	179	.200*	0.994	179	0.624	0.822	0.222
Nasal vertical prominence (tr-prn-gn)	0.033	179	.200*	0.995	179	0.861	0.373	-1.026
Transverse nasal prominence (zy(l)-prn-zy(r))	0.046	179	.200*	0.987	179	0.084	1.747	-0.668
Nasolabial angle (prn-sn-ls)	0.056	179	.200*	0.988	179	0.125	-2.026	-0.049
Nasofrontal angle (g-n-prn)	0.048	179	.200*	0.991	179	0.33	-1.079	0.846
Forehead nasal angle (tr-n-prn)	0.053	179	.200*	0.989	179	0.203	-1.195	1.121
Chin prominence (go(l)-gn-go(r))	0.071	179	0.029	0.984	179	0.038	2.389	0.754
Transverse nasal prominence 2 (t(l)-prn-t(r))	0.06	179	.200*	0.984	179	0.036	2.384	-0.339
Nasion depth angle (zy(l)-n-zy(r))	0.053	179	.200*	0.992	179	0.408	0.213	1.515
Nasomental angle (n-prn-pg)	0.044	179	.200*	0.997	179	0.969	-0.19	-0.482

This study identified potential outliers, represented by extreme values in the dataset and caused by either wrong assignment of facial landmarks and subsequent wrong calculations of the distances using an Excel spreadsheet. The errors in assigning the landmarks in a few images were found in two main landmark pairs: the left and right zygion and gonion. The location of all landmarks was always performed in the same order: from number 1 to number 32. For example, the left gonion being manually appointed number 9, the right gonion number 10, the right zygion – number 26 and the left zygion – number 27 (Table 4 and Figure 29). In a few images this order was disrupted because of human error. For example, the right gonion was allocated first and incorrectly assigned as number 9 and then the left gonion as number 10, whereas instead, it should be in the opposite order. In addition, when some of the landmarks had to be edited or deleted as a part of the annotation process, the Geomagic software arranged these updated landmarks in a new order, causing errors in the Excel formulas that were based on the specific order of these values. Conducting the test for normality of the measurements, assessed the normal distribution of the data and found the extreme outliers (potentially representing errors in the measurements), which were corrected. The final dataset was subsequently processed for statistical analysis of potential associations with genetic markers (as detailed in Chapter 5).

4.5.4. Variation in craniofacial measurements between ethnic groups

The purpose of this study was to test several craniofacial distances for normal distribution in six populations as well as for sex and ethnic variation and to compare these data with published resources.

Despite the abnormal distribution of several direct craniofacial measurements in the original (complete) dataset, these same measurements showed normal distribution following splitting data by ethnic origin (Table 28).

The distances were analysed for the distribution of mean values between four main populations: Asian, African, Caucasian and Indian (other populations were not analysed due to insufficient sample size). The resulting box plots (data not shown) demonstrated a very wide distribution of data in the Caucasian population, with high standard

deviation and numerous outliers, where outliers were defined as either ≥3x interquartile range (IQR) above the third quartile or ≥3×IQR below the first quartile.

As a result, the Caucasian group was split into three relatively homogeneous sub-populations as East Europeans (EE), West Europeans (WE) and Middle Eastern (ME). The results that compared these six population groups for direct craniofacial measurements and Cephalic index in both sexes are shown in Table 28 and Figure 48. For the head height and length, the results were consistent with published data [17, 28]. Males in general, demonstrated longer mean distances than females for all direct measurements. Head height was found to be the highest in the African population and the shortest in the Indian population in both sexes. Head width was the widest in the African population in females and in the Asian population in males, while being the narrowest in the ME population in females and in the African population in males. Head length was found to be the longest in the African population and the shortest in the Asian population in both genders, which is similar to published data [17, 18]. The Cephalic index showed a distribution from 77.4 to 84.5 in females and between 74.7 and 85.6 in males, with higher values in the Asian population in both genders, demonstrating a brachycephalic or even a hyperbrachycephalic head.

Among the population groups tested, Asians were found to have the widest, but the shortest head. Africans had the longest head as Asians, especially in females; while Indians and three Caucasian sub-populations demonstrated various, but generally medial mean values for all manual measurements (Figure 48). These results were in general consensus with previous published data.

All linear and angular distances demonstrated normal distribution in all ethnic groups tested (only partial data is shown due to space limits). Ten linear and six angular distances were summarized for the purpose of this study (Tables 29-31 and Figures 49-51). All five horizontal linear distances tested were longer in males, except for en-en distance in Asians, which was slightly longer in females (Table 29 and Figure 50). Two vertical distances (n-prn and tr-gn) were longer in males, except for the tr-gn distance in the African population, which was longer in females (Table 29 and Figure 49). The other three vertical distances demonstrated no clear pattern. Two angular distances (transverse nasal prominence and nasal tip prominence) showed higher angles for females, while four angles were represented by a mixed pattern (Table 31 and Figure 51).

To summarize, males in general, had longer and wider faces, with more prominent larger noses and more widely set eyes in most populations tested. Population wise, the

Asians demonstrated the widest face and cranium, the Africans the widest nose, lips and binocular width, while the Middle Eastern males and Indian females were found to have the longest noses. These results are in general consensus with published data and may form foundation for extended follow up study with a larger sample set.

Table 28. Average distances, standard errors and Shapiro-Wilk test generated p-values of three direct craniofacial measurements and cephalic index in various population groups tested.

Descriptives - females					Descriptives - males				
	Population	Mean distance	Std. Error	Shapiro-Wilk p-value		Population	Mean distance	Std. Error	Shapiro-Wilk p-value
V-Gn (Head height) mm	African	247.939	1.582	0.663	V-Gn (Head height) mm	African	262.244	3.033	0.631
	Asian	240.581	1.206	0.115		Asian	254.019	2.214	0.587
	EE	239.511	1.314	0.034		EE	252.824	2.432	0.993
	Indian	237.237	1.191	0.010		Indian	246.901	1.965	0.857
	ME	240.393	2.560	0.724		ME	253.584	1.568	0.925
	WE	240.102	0.604	0.027		WE	256.623	0.787	0.089
Eu-Eu (Head width) mm	African	157.279	3.035	0.498	Eu-Eu (Head width) mm	African	152.448	5.977	0.936
	Asian	153.623	1.145	0.619		Asian	161.995	1.686	0.759
	EE	149.576	1.531	0.310		EE	155.201	1.816	0.520
	Indian	148.482	1.115	0.215		Indian	154.586	1.106	0.425
	ME	146.677	1.839	0.996		ME	155.575	1.350	0.918
	WE	148.334	0.362	0.631		WE	155.321	0.446	0.442
G-Op (Head length) mm	African	199.627	3.010	0.435	G-Op (Head length) mm	African	203.734	1.974	0.761
	Asian	182.059	1.195	0.783		Asian	189.251	1.708	0.590
	EE	188.538	1.195	0.946		EE	197.233	4.014	0.710
	Indian	183.821	1.069	0.530		Indian	194.788	1.901	0.639
	ME	187.599	1.941	0.990		ME	193.742	1.225	0.211
	WE	191.804	0.456	0.304		WE	202.854	0.504	0.682
Cephalic Index	African	78.896	1.937	0.857	Cephalic Index	African	74.741	2.224	0.956
	Asian	84.491	0.869	0.427		Asian	85.596	0.407	0.514
	EE	79.400	1.046	0.392		EE	78.913	2.016	0.289
	Indian	80.829	0.735	0.079		Indian	79.513	0.994	0.724
	ME	78.218	0.877	0.174		ME	80.348	0.753	0.294
	WE	77.417	0.257	0.170		WE	76.597	0.238	0.904

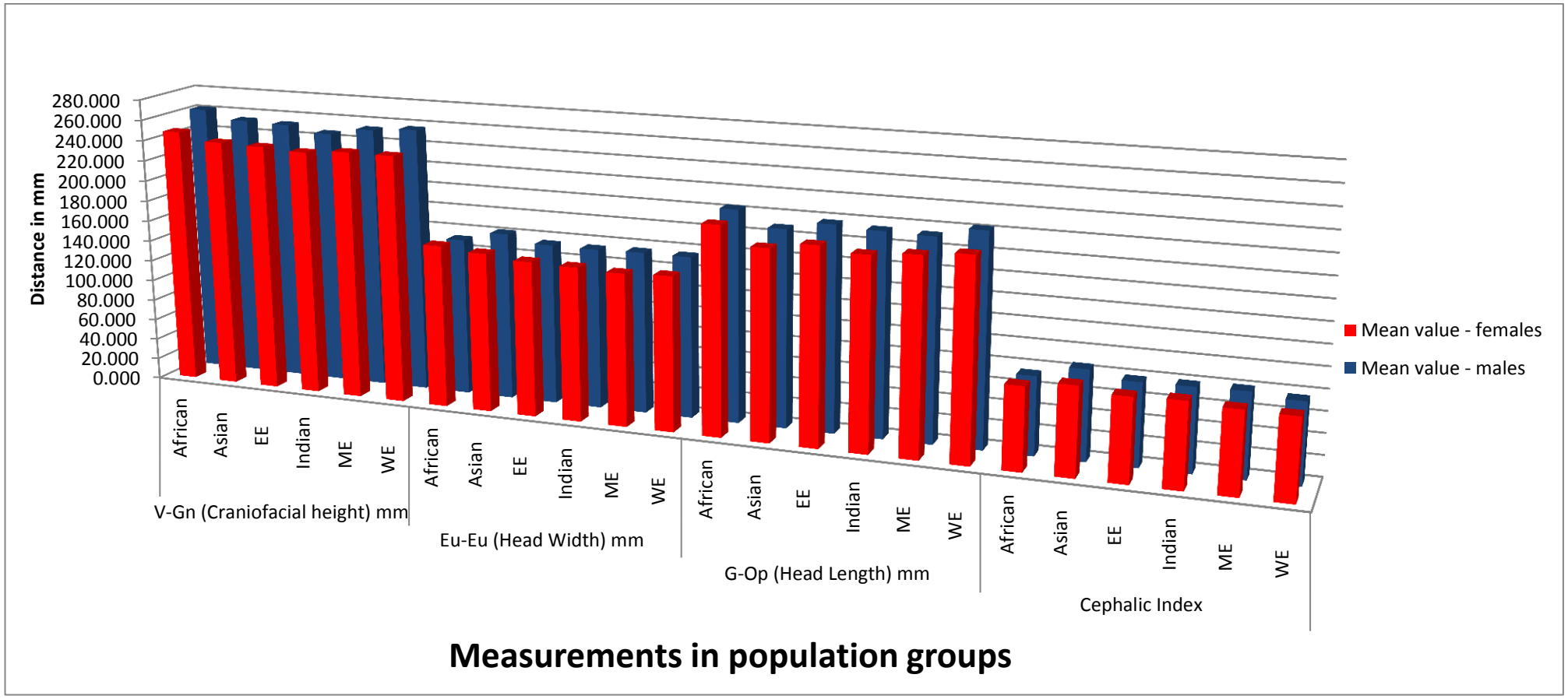


Figure 48. Comparison of ethnic and sex –related variation in direct craniofacial measurements in various population groups.

Table 29. Average distances, standard errors and Shapiro-Wilk test generated p-values of five linear vertical facial distances in various population groups tested.

		Descriptives - females				Descriptives - males			
Population		Mean value - females	Std. Error	Shapiro-Wilk p-value		Population	Mean value - males	Std. Error	Shapiro-Wilk p-value
tr-gn	African	187.164	1.901	0.399		African	181.824	8.041	0.876
	Asian	179.393	2.217	0.129		Asian	188.439	2.855	0.170
	EE	176.365	1.619	0.874		EE	182.078	4.082	0.579
	Indian	174.327	1.954	0.593		Indian	175.984	1.723	0.995
	ME	174.551	3.385	0.063		ME	183.000	1.881	0.451
	WE	178.670	0.667	0.752		WE	186.651	1.018	0.312
tr-n	African	83.093	3.910	0.551		African	69.762	4.203	0.178
	Asian	73.729	1.820	0.750		Asian	75.132	2.131	0.066
	EE	72.425	1.407	0.287		EE	68.385	5.264	0.127
	Indian	70.714	1.494	0.745		Indian	64.345	1.709	0.831
	ME	69.316	2.080	0.459		ME	66.393	1.419	0.853
	WE	73.168	0.555	0.075		WE	71.748	0.927	0.129
n-prn	African	37.374	2.450	0.411		African	43.377	2.457	0.420
	Asian	42.361	0.655	0.064		Asian	45.455	0.883	0.330
	EE	42.893	0.751	0.387		EE	46.445	1.033	0.987
	Indian	44.083	0.940	0.553		Indian	45.474	0.827	0.484
	ME	41.788	1.051	0.311		ME	49.864	0.917	0.897
	WE	42.924	0.335	0.555		WE	46.401	0.393	0.072
ps(l)-pi(l)	African	8.442	0.220	0.653		African	8.722	0.993	0.155
	Asian	8.769	0.334	0.959		Asian	8.622	0.582	0.492
	EE	8.819	0.325	0.885		EE	9.337	0.897	0.302
	Indian	11.105	0.435	0.665		Indian	9.524	0.310	0.317
	ME	9.580	0.427	0.288		ME	9.008	0.319	0.174
	WE	9.389	0.131	0.337		WE	8.669	0.159	0.063
li-sto	African	12.640	1.106	0.160		African	12.231	0.375	0.131
	Asian	9.861	0.356	0.453		Asian	10.070	0.923	0.214
	EE	8.597	0.349	0.977		EE	8.205	0.327	0.180
	Indian	10.433	0.436	0.427		Indian	10.184	0.381	0.707
	ME	10.168	0.937	0.684		ME	10.523	0.475	0.792
	WE	9.081	0.147	0.865		WE	8.823	0.219	0.069

Table 30. Average distances, standard errors and Shapiro-Wilk test generated p-values of five linear horizonthal facial distances in various population groups tested.

Descriptives - females						Descriptives - males			
Population		Mean value - females	Std. Error	Shapiro-Wilk p-value		Population	Mean value - males	Std. Error	Shapiro-Wilk p-value
zy-zy	African	137.973	1.879	0.379		African	142.194	7.082	0.812
	Asian	142.130	1.543	0.020		Asian	147.781	1.650	0.613
	EE	137.686	1.479	0.332		EE	142.049	2.056	0.297
	Indian	134.855	1.551	0.188		Indian	141.938	1.856	0.026
	ME	131.970	1.261	0.858		ME	141.752	1.823	0.065
	WE	135.586	0.476	0.652		WE	141.406	0.723	0.029
go-go	African	119.734	2.233	0.488		African	128.499	4.374	0.083
	Asian	125.931	1.850	0.476		Asian	131.214	1.572	0.036
	EE	120.549	1.517	0.394		EE	128.158	1.869	0.727
	Indian	117.599	1.983	0.866		Indian	129.309	1.762	0.020
	ME	115.133	2.041	0.435		ME	127.098	1.942	0.182
	WE	118.734	0.558	0.638		WE	127.673	0.790	0.161
al-al	African	39.647	2.171	0.634		African	46.524	1.101	0.654
	Asian	36.987	0.583	0.887		Asian	39.559	0.666	0.804
	EE	34.477	0.501	0.121		EE	36.990	0.810	0.761
	Indian	35.052	0.655	0.993		Indian	38.578	0.858	0.003
	ME	32.598	1.144	0.729		ME	37.454	0.515	0.903
	WE	32.807	0.201	0.860		WE	36.332	0.297	0.021
en-en	African	38.927	1.349	0.304		African	41.139	3.222	0.526
	Asian	39.248	0.551	0.166		Asian	38.181	0.625	0.343
	EE	37.391	0.823	0.985		EE	39.489		0.652
	Indian	33.925	0.752	0.492		Indian	35.956	0.731	0.836
	ME	34.200	1.324	0.063		ME	36.737	0.585	0.561
	WE	35.521	0.232	0.182		WE	37.534	0.363	0.735
ex-ex	African	89.580	1.260	0.095		African	92.875	3.728	0.376
	Asian	86.192	1.055	0.858		Asian	89.247	0.872	0.754
	EE	85.014	0.875	0.245		EE	88.207	1.460	0.434
	Indian	87.548	0.968	0.209		Indian	89.057	0.853	0.971
	ME	81.894	2.098	0.853		ME	87.758	1.177	0.930
	WE	84.213	0.322	0.592		WE	86.876	0.423	0.290

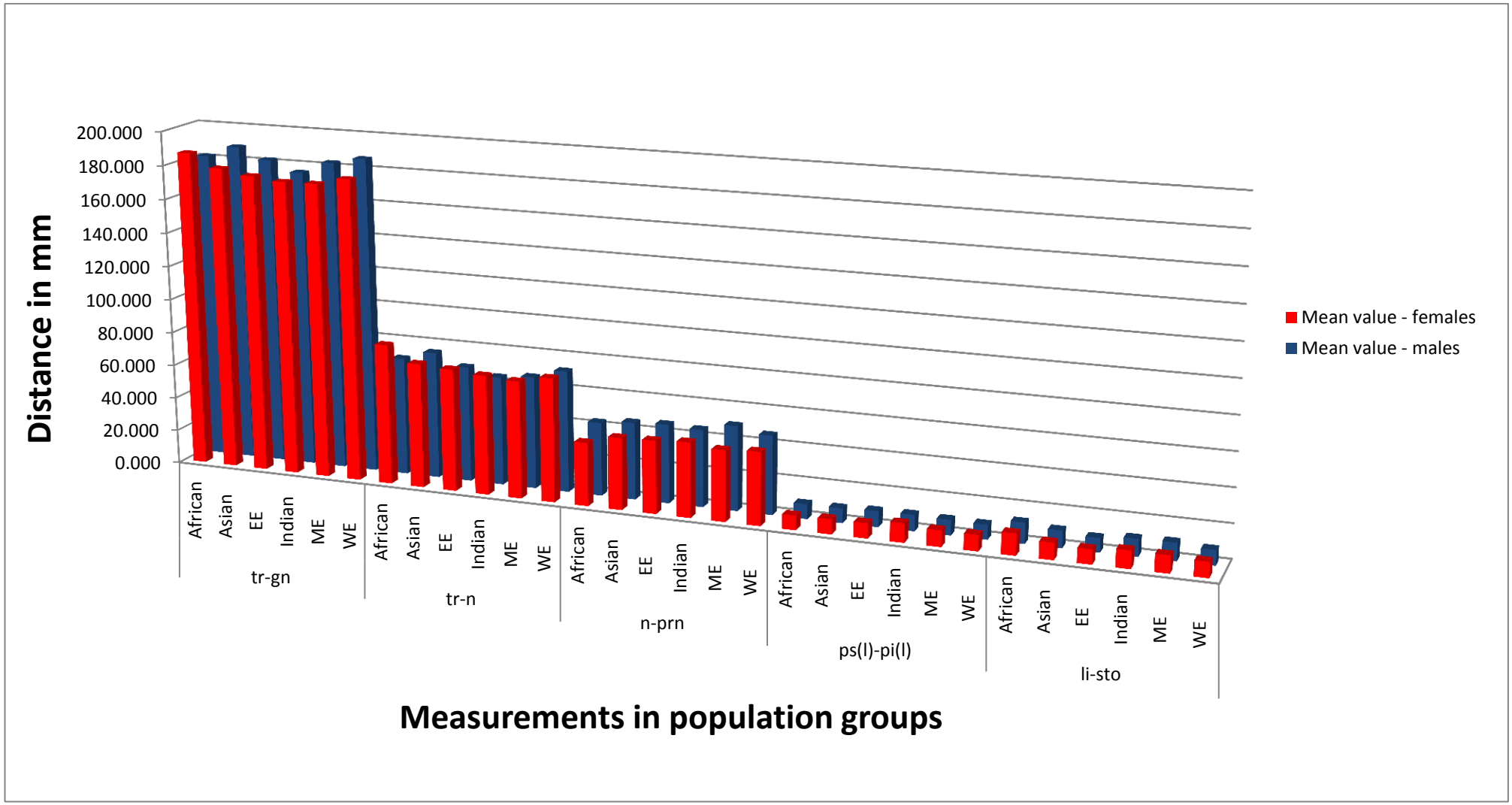


Figure 49. Comparison of ethnic and sex –related variation in linear vertical facial measurements in various population groups.

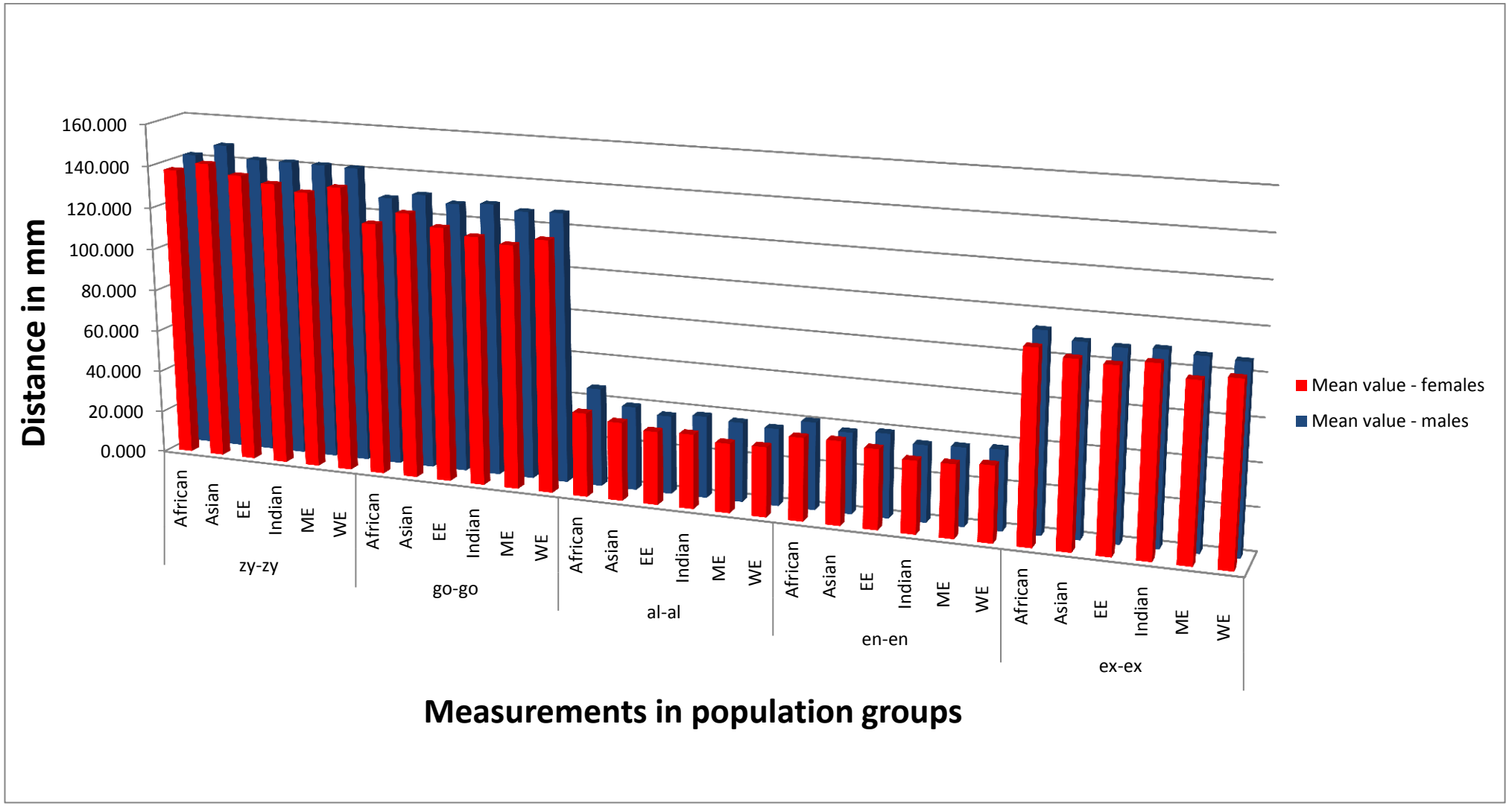


Figure 50. Comparison of ethnic and sex –related variation in linear horizontal facial measurements in various population groups.

Table 31. Average distances, standard errors and Shapiro-Wilk test generated p-values of six angular facial distances in various population groups tested.

		Descriptives - females				Descriptives - males		
	Population	Statistic	Std. Error	Shapiro-Wilk p-value	Population	Statistic	Std. Error	Shapiro-Wilk p-value
Nasal tip angle (n-prn-sn)	African	102.368	2.115	.768	African	100.189	2.550	.679
	Asian	101.299	1.370	.790	Asian	99.061	1.506	.340
	EU	94.284	1.297	.471	EU	93.646	1.779	.435
	Indian	92.517	1.213	.791	Indian	95.431	1.139	.804
	ME	97.490	2.464	.589	ME	92.984	1.037	.217
	WE	94.889	0.391	.667	WE	93.527	0.522	.512
Transverse nasal prominence (zy(l)-prn-zy(r))	African	87.876	1.704	.369	African	85.975	2.219	.381
	Asian	89.773	0.881	.129	Asian	89.135	0.696	.258
	EU	81.958	1.088	.588	EU	79.706	1.422	.628
	Indian	82.253	0.787	.259	Indian	80.732	0.674	.136
	ME	82.542	1.211	.167	ME	80.531	0.880	.043
	WE	82.241	0.285	.879	WE	80.395	0.398	.956
Nasolabial angle (prn-sn-ls)	African	105.143	3.603	.442	African	99.244	2.345	.352
	Asian	112.619	1.776	.472	Asian	115.747	2.654	.426
	EU	118.786	2.174	.489	EU	125.511	2.147	.509
	Indian	115.557	2.383	.906	Indian	120.307	2.026	.212
	ME	121.135	3.947	.389	ME	120.591	2.291	.415
	WE	121.572	0.614	.721	WE	121.989	0.929	.169
Nasofrontal angle (g-n-prn)	African	146.359	0.759	.931	African	142.569	1.711	.979
	Asian	149.527	0.999	.639	Asian	149.320	1.148	.794
	EU	150.500	1.341	.339	EU	148.387	2.533	.314
	Indian	150.387	1.035	.881	Indian	151.681	1.196	.397
	ME	151.540	1.469	.578	ME	154.441	1.209	.714
	WE	149.276	0.370	.712	WE	150.112	0.550	.421
Nasion depth angle (zy(l)-n-zy(r))	African	105.844	1.302	.562	African	99.238	1.923	.196
	Asian	108.047	1.112	.047	Asian	109.622	0.982	.200
	EU	99.683	1.405	.184	EU	97.334	2.410	.340
	Indian	102.531	0.647	.603	Indian	97.680	1.007	.597
	ME	99.646	1.763	.615	ME	96.979	1.054	.600
	WE	99.831	0.296	.848	WE	97.028	0.494	.408
Nasomental angle (n-prn-pg)	African	132.433	2.209	.045	African	134.588	2.386	.156
	Asian	135.088	0.777	.108	Asian	132.399	1.006	.882
	EU	129.236	1.136	.072	EU	127.386	1.287	.086
	Indian	128.280	1.241	.220	Indian	129.334	1.170	.271
	ME	129.329	1.452	.500	ME	130.147	1.174	.959
	WE	129.765	0.353	.381	WE	130.731	0.485	.663

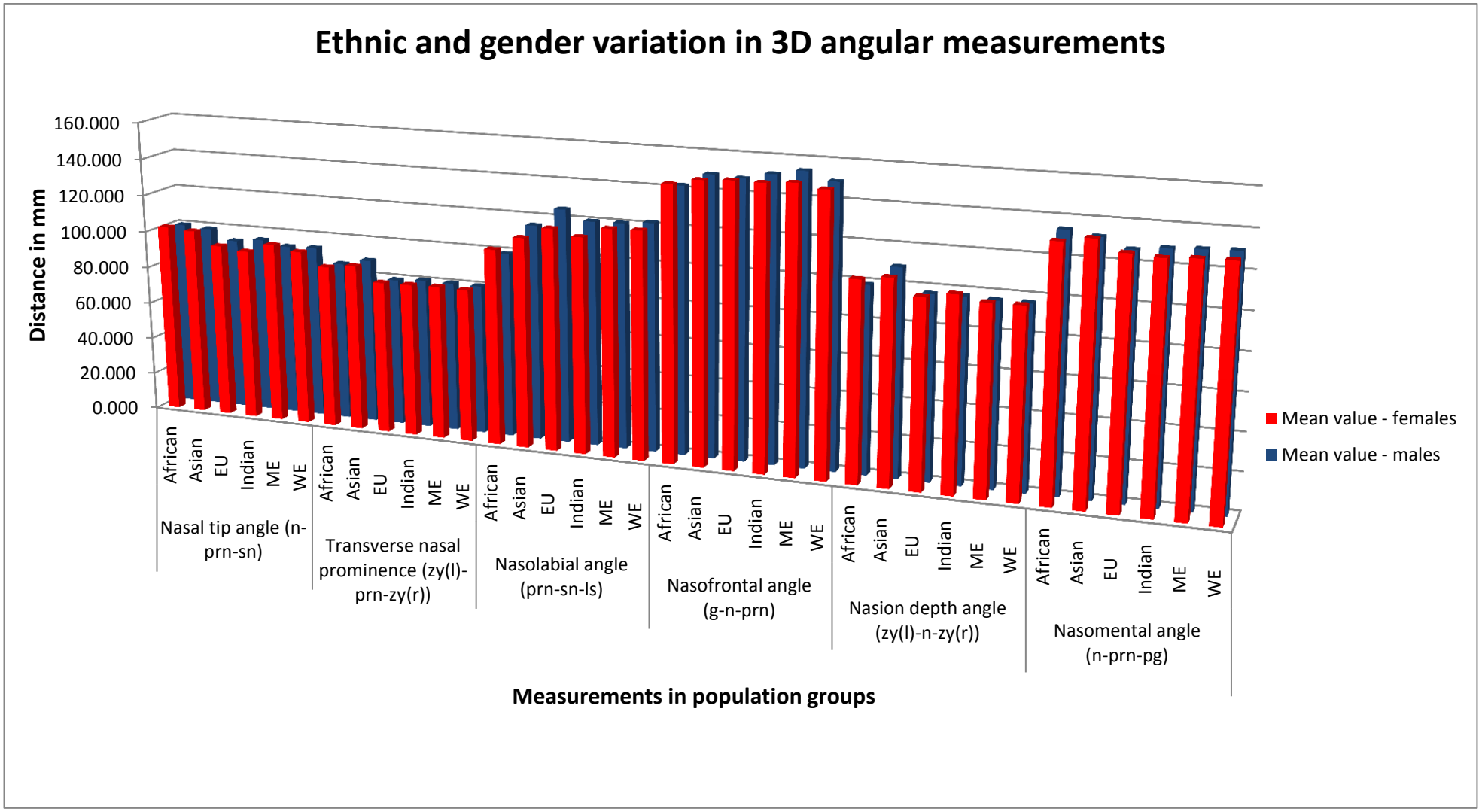


Figure 51. Comparison of ethnic and sex –related variation in the angular facial distances in various population groups.

4.6. Principal component analysis

The main goal of principal component analysis (PCA) was to reduce the number of traits for subsequent association study. Principal component analysis of all 92 measurements (detailed in Sections 2.9.2, 2.9.3 and 2.9.4) as one set, produced 20 principal components, which accounted for almost 90% of the variance (Table 32 and Figure 52). The efficiency of PCA can be demonstrated by the fact that the first three components and another seven components explained 1/3 and 2/3 of the total variance respectively. These results demonstrated that subsequent analysis of genotype association may be performed with 20 (or less) principal components (instead of 92 individual measurements), which can potentially provide information on approximately 90% of the craniofacial traits used.

However, given that direct measurements were tested together with respective ratios, this analysis did not produce a clear anthropometric pattern of linear distances (Table 33). As a result, the linear and angular measurements were split from ratios and analysed separately (Tables 34 and 35).

Table 32. Rotated component matrix results for linear and angular measurements, including ratios between these measurements.

All the measurements - Total Variance Explained

Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	13.379	14.542	14.542
2	9.238	10.041	24.584
3	7.584	8.244	32.828
4	6.266	6.811	39.638
5	5.480	5.957	45.595
6	5.118	5.563	51.159
7	3.844	4.179	55.337
8	3.752	4.078	59.416
9	3.621	3.936	63.352
10	3.444	3.744	67.096
11	3.258	3.541	70.637
12	2.862	3.111	73.748
13	2.644	2.874	76.622
14	2.094	2.276	78.898
15	1.829	1.988	80.886
16	1.826	1.985	82.872
17	1.729	1.880	84.751
18	1.668	1.813	86.564
19	1.579	1.716	88.280
20	1.378	1.498	89.778

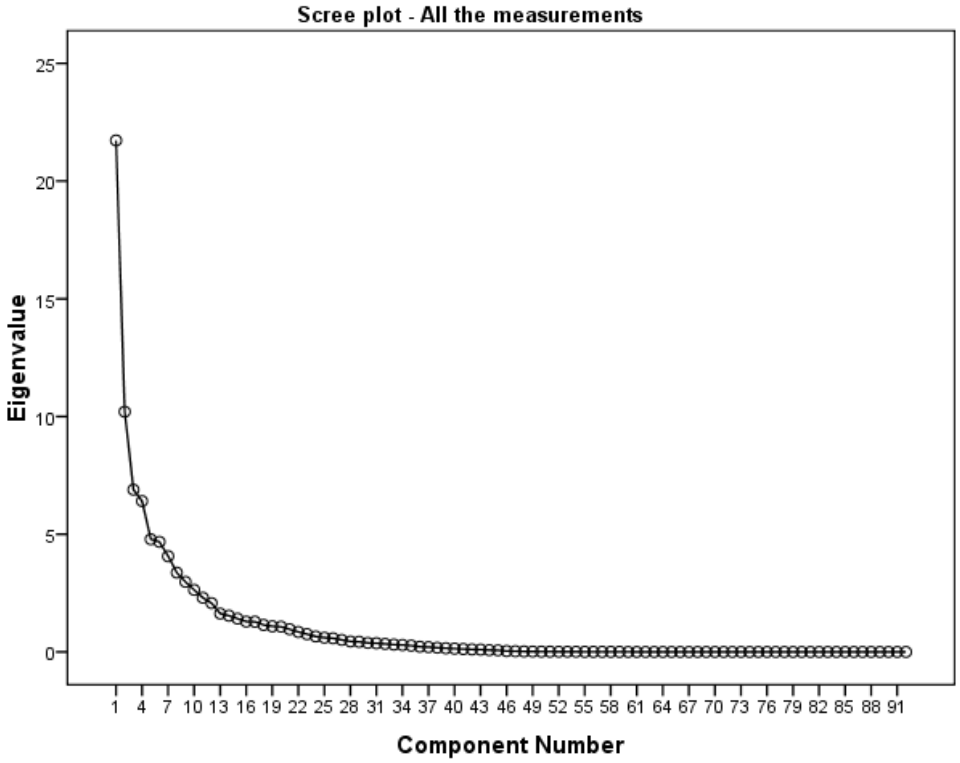


Figure 52. The scree plot of linear and angular measurements, including ratios between these measurements.

Table 33. Total variance in linear and angular measurements, including ratios explained by principal components.

An eigenvalue threshold of 0.6 was applied in order to produce a clearer pattern of principal components.

All the measurements - Rotated Component Matrix

	Component																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
t(l)-g	.905																			
t(l)-n	.870																			
t(r)-tr	.843																			
t(l)-prn	.840																			
t(l)-tr	.833																			
t(l)-gn	.786																			
t(r)-t(l)	.773																			
t(r)-gn	.751																			
Eu-Eu (Head Width) mm	.654																			
G-Op (Head Length) mm	.640																			
V-Gn (Craniofacial height) mm	.606																			
zy-zy																				
sn-gn		.901																		
Mandible index: (sto-gn)*100 /(go-go)		.875																		

sto-gn		.867																					
Lower face height ratio (sn-gn*100/go-go)		.867																					
prn-gn		.843																					
n-gn		.727																					
Upper face height ratio (n-sn*100/sn-gn)		-.700	.676																				
Anterior face height 2 ratio (n-gn*100/zy-zy)		.686																					
sl-gn		.676																					
Anterior face height 1 (n-gn*100/go-go)		.646																					
g-gn																							
g-pg																							
Upper face height ratio (n-sn*100/zy-zy)			.918																				
n-sn			.889																				
Upper face height ratio (n-sn*100/go(r)-go(l))			.837																				
n-prn			.829																				
Nose-face height index n-sn/n-gn		-.674	.696																				
n-sto			.676																				
tr-n				.933																			
g-tr				.931																			
Tr-g*100/sn-gn				.833																			
Forehead height ratio (tr-n*100/go(r)-go(l))				.831																			
Face height index: (n-gn)*100/(tr-gn)				-.727																			
tr-gn				.683																			
Total anterior face height ratio (tr-gn*100/zy-zy)				.622																			
go(r)-tr				.612																			
go(l)-tr																							
Forehead nasal angle (tr-n-prn)																							
n-zy(l)					.802																		
zy(r)-al(r)					.780																		
zy(l)-al(l)					.772																		
n-zy(r)					.759																		
Nasion depth angle (zy(l)-n-zy(r))					-.601																		
Transverse nasal prominence (zy(l)-prn-zy(r))																							
Mandible – Face width ratio (go-go*100/(zy-zy)						.895																	
go-go						.746																	
go-go*100/ex-ex						.731																	

gn-go(l)						.710															
gn-go(r)						.707															
prn-go(r)																					
prn-go(l)																					
Nose-face width index al-al/zy-zy							.923														
al-al							.846														
Nasal Index al-al/n-sn			-.650				.700														
Interendocanthion distance ratio (en-en*100/al-al)							-.683				.641										
prn-al(l)																					
prn-al(r)																					
go(l)-zy(l)								.910													
go(r)-zy(r)								.893													
tr-zy(r)																					
zy(l)-gn																					
tr-zy(l)																					
zy(r)-gn																					
Cephalic Index									.861												
Head width – Craniofacial height ratio (eu-eu*100/v-gn)									.661												
Transverse nasal prominence 2 (t(l)-prn-t(r))																					
ps(l)-pi(l)										.798											
ps(r)-pi(r)										.786											
en(l)-ex(l)										.735											
en(r)-ex(r)										.648											
ex-ex										.629											
Mouth width ratio (ch-ch*100/en-en)											-.918										
en-en											.801										
Intercanthal index en(R)-en(L)*100/ex(R)-ex(L)											.744										
sn-prn												.895									
Nasal Tip Protrusion –Nose height index: (sn-prn)x100/(n-sn)												.893									
Nasal tip prostrusion - width index sn-prn/al-al												.765									
Nasomental angle (n-prn-pg)													.869								
Nasal vertical prominence (tr-prn-gn)													.733								
ls-sto														.826							
li-sto														.768							
Chin prominence (go(l)-gn-go(r))															.692						
ch-ch																					
v-gn*100/height																.727					

Nasofrontal angle (g-n-prn)																.634			
Nasolabial angle (prn-sn-ls)																	.738		
Nasal tip angle (n-prn-sn)																	.732		
t(l)-pa(l)																		.663	
sa(l)-sba(l)																			
ex(R)-en(R)*100/en(L)-ex(L)																			.958

Tables 34, 35 and Figure 53 demonstrate the results of the PCA on the linear and angular distances, not including the ratios.

Table 34. Total variance in linear and angular measurements explained by principal components.

Total Variance Explained – Linear and angular distances			
Component	Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	6.762	12.523	12.523
2	5.947	11.014	23.536
3	4.738	8.773	32.310
4	4.440	8.222	40.531
5	4.046	7.492	48.023
6	3.684	6.822	54.845
7	3.422	6.336	61.181
8	2.779	5.147	66.328
9	2.550	4.722	71.049
10	2.227	4.124	75.174
11	1.889	3.498	78.672
12	1.685	3.120	81.792
13	1.612	2.985	84.776

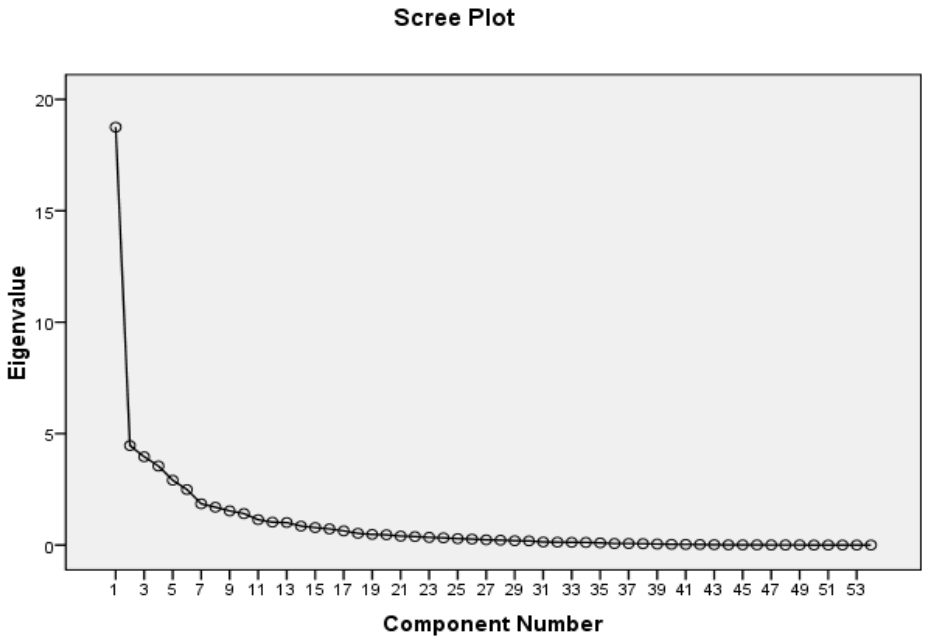


Figure 53. The scree plot of PCA performed on the linear and angular measurements, excluding ratios.

The results of PCA on linear and angular measurements are summarised in Table 34 and Figure 53. Thirteen principal components for all 54 distances were obtained, while three and eight principal components explained 1/3 and 2/3 of the total variance respectively (Table 33).

Separating direct craniofacial measurements (linear and angular) from ratios, and performing PCA on each group provided a more explanatory picture (in biological sense) of principal components. Most of the distances in each principal component were automatically grouped according to their similarity in Euclidean space orientation and corresponding landmarks. For easier presentation of the data and due to space limits, only the linear distances are shown (Table 35). For example, the first principal component was formed by horizontal distances, which all include the tragion (located on the ear). The second principal component represented vertical distances, which all include the gnation or pogonion (chin area). The third principal component was represented by horizontal measurements that all included the zygion. The fourth principal component incorporated horizontal measurements, which all included the gonion as one of the paired landmarks. The same rational is applicable for the rest of the principal components. Two distances ex-ex and ch-ch were not included in any principal component. This exclusion was most likely due to inaccurate measurements of

both eye and lip areas, as a result of poor laser coverage, as previously discussed in Section 4.4.

Figure 54 illustrates an example of the linear craniofacial distances, coloured according to the principal components they represent (as detailed in Table 35).

Table 35. Rotated Component Matrix results for linear craniofacial measurements.

Rotated Component Matrix - Linear and angular distances

		Component												
		1	2	3	4	5	6	7	8	9	10	11	12	13
PC I	t(l)-g	.852												
	t(l)-n	.802												
	t(l)-tr	.762												
	t(r)-tr	.744												
	t(l)-prn	.714												
	t(r)-t(l)	.629												
	t(l)-gn	.619												
	t(r)-gn	.541												
PC II	sn-gn		.904											
	sto-gn		.900											
	prn-gn		.866											
	sl-gn		.765											
PC II + VI	n-gn		.740				.542							
PC II	g-gn		.621											
	g-pg		.530											
PC III	zy(l)-al(l)			.846										
	zy(r)-al(r)			.836										
	n-zy(l)			.796										
	n-zy(r)			.761										
	zy-zy			.620										
PC IV	gn-go(l)				.851									
	gn-go(r)				.844									
	go-go				.698									
	prn-go(r)				.681									
	prn-go(l)				.664									
PC V	g-tr					.901								
	tr-n					.875								
	tr-gn					.729								

	go(r)-tr go(l)-tr					.713 .687									
PC V + VII	tr-zy(l)					.532		-.505							
PC VI	n-prn n-sn n-sto						.902 .882 .767								
PC VII	go(l)-zy(l) go(r)-zy(r)							.924 .916							
PC IV +VII	tr-zy(r)					.526		-.567							
PC III + VII	zy(l)-gn zy(r)-gn			.514 .500				.523 .503							
PC VIII	ps(r)-pi(r) ps(l)-pi(l) en(l)-ex(l) en(r)-ex(r)								.871 .869 .688 .628						
PC IX	al-al									.734					
PC X	sn-prn prn-al(l) prn-al(r)										.825 .655 .641				
PC XI	li-sto ls-sto											.813 .778			
PC XII	t(l)-pa(l) sa(l)-sba(l)												.855 .688		
PC XIII	en-en													.545	

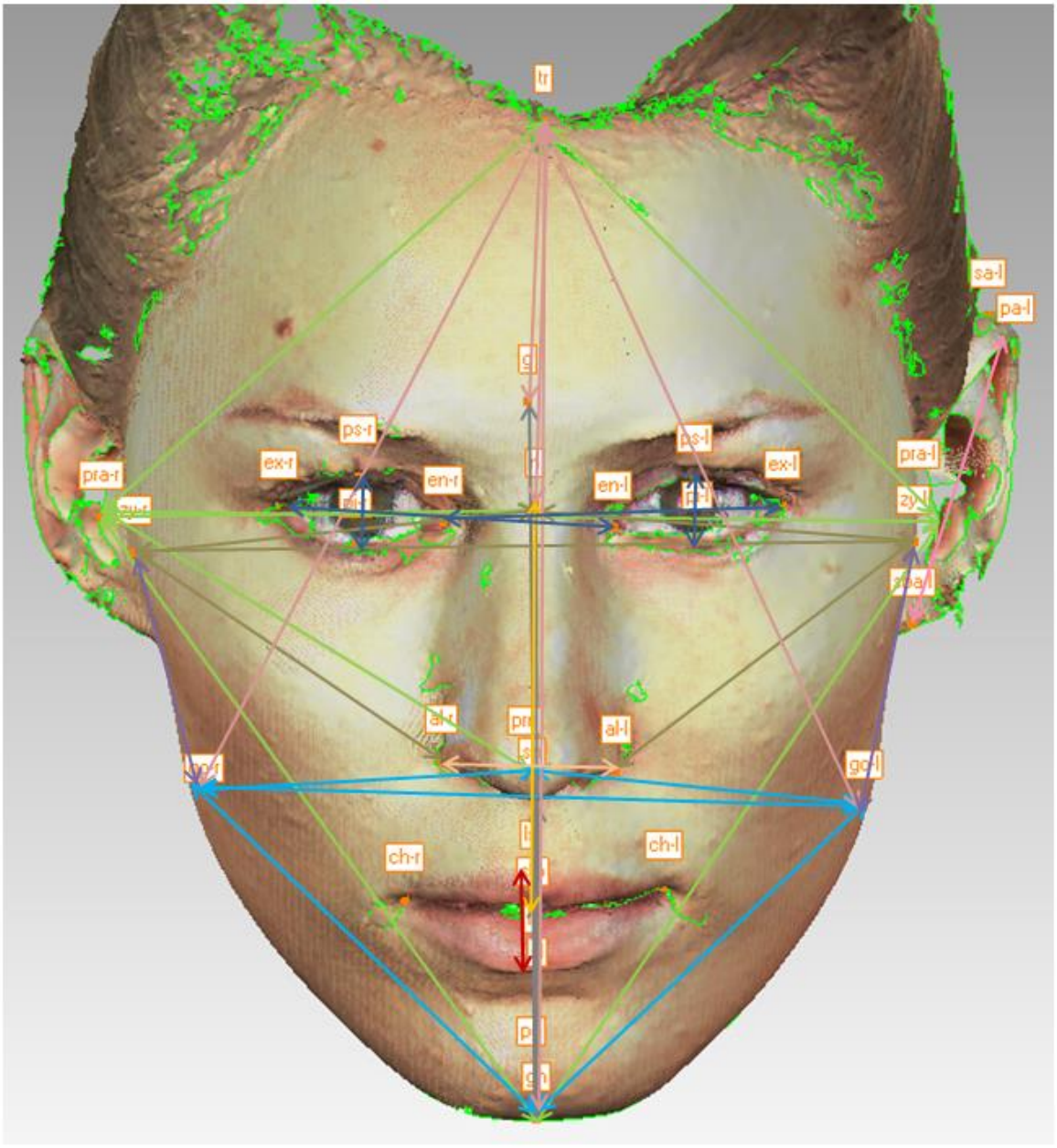


Figure 54. An illustration of the linear craniofacial distances according to respective colouring of the major principal components. A few distances are not shown due to image resolution.

These results supply further confirmation of the validity of the 3D craniofacial measurements and provide a solid foundation for the subsequent genetic association analysis described in Chapter 5.

Chapter 5

Association and prediction study of the externally visible traits and ancestry

5.1. Introduction

The overall goal of this project was to contribute to prediction of Externally Visible Traits (EVTs), such as facial appearance, eye, skin and hair pigmentation and also ancestry. The main aim of this study was to find potential associations between candidate genetic markers and craniofacial traits. A secondary aim was to analyse associations of observed pigmentation traits (eye, skin and hair colour) with previously published or novel markers as well as with main ancestry groups. A further aim was to assess the potential of a set of significantly associated SNPs in establishing prediction models of EVT and ancestry.

The association and prediction analyses of pigmentation traits and ancestry origin were important to investigate for two reasons. Firstly, conducting an association study with previously associated markers allowed confirmation of the validity of the statistical modelling, thus ensuring results for novel (craniofacial) traits are robust. Secondly, performing prediction modelling provides an opportunity to test the predictive power of the pigmentation and ancestry SNPs set used in this study, potentially improving the current forensic assays.

The following section summarizes results of the genetic association analyses for ancestry, pigmentation and craniofacial phenotypes as well as evaluation of statistical models for these phenotypes prediction.

5.1.1. Ancestry informative markers (AIMs)

Ancestry Informative markers (AIMs) are generally defined as genetic markers that display low heterozygosity and high Fst values [247]. In the genetic context, these markers can be found in various genes that have been subject to positive selection under the adaptation process, such as pigmentation–related or immunity–regulating genes [370]. As a result, AIMs may provide indirect information on phenotype. For example, in North Europeans, skin pigmentation is lighter compared to most other populations, however this correlation cannot be used as a precise predictor of the phenotype and must be treated accordingly [371].

In general, ancestry prediction focuses on three main categories or "dimensions" [372, 373]. The first is the *continental ancestry*, which assumes the existence of approximately five major populations that originally gave rise to existing major

population groups at approximately 100,000 years ago. The second is the *biogeographic ancestry*, which relates to the geographic location of a specific population living in this area and usually refers to unlinked autosomal ancestry informative markers (AIMs). The third category is the *lineage ancestry*, which refers to the paternal or maternal family history. This history can be revealed through genotyping of the uniparental markers (mt-DNA or Y-chromosome haplotypes). In the forensic context, prediction of the biogeographic and lineage ancestries can be used as intelligence information to solve crimes with no suspect available and assist in missing person and DVI cases. In general, the more geographically localized the information that can be predicted is, the more useful it is for the investigative purpose. However, estimation of genetic ancestry can be challenging, given that human demographic history and specifically population admixture as a result of recent migration dynamics. This project explored only the biogeographical ancestry associations and prediction queries.

Estimating an individual's ancestry from DNA markers is not a novel concept [249]. Research on ancestry prediction using genetic markers has rapidly progressed from a continental resolution down to a more focused geographical location, forming the basis for forensically–relevant assays [374, 375]. Some of these assays are able to predict additional traits by incorporating pigmentation-informative markers, such as skin, hair or eye colour from DNA samples of various amount and quality [9, 16, 267, 291, 376-378]. A few ancestry-predictive assays have been recently validated as successful investigative tools in several forensic cases [16, 379]. For example, in the Louisiana serial killer case, the police were searching for a Caucasian suspect (based on eyewitnesses testimony), when in fact he was predicted to be 85% sub-Saharan African and 15% Native American, as revealed from ancestry DNA testing and subsequently matching the ancestry origin of the offender [16]. In addition, the use of genetic approaches for predicting ancestry of human remains is highly preferable over forensic anthropology, with the latter considered unreliable and prone to errors [380]. Nevertheless, it should be emphasized that ancestry prediction using a genetic approach should be treated as an investigative lead and not as an identity-confirmatory test.

The potential correlation between AIMs and SNPs involved in the craniofacial traits variation has not yet been explored. Given that ancestry is correlative to external appearance and that the markers for this project were chosen with high Fst values, an overlap between ancestry, pigmentation and craniofacial markers is expected. Therefore, the incorporation of such markers in an assay may simultaneously provide investigative

information on the facial appearance, pigmentation and bio-geographical ancestry of an individual.

5.1.2. Pigmentation traits in humans

Eye, skin and hair colour are among the most visible external features. Knowledge of the genetics of human pigmentation has been enriched by studies of mice and zebra fish mutants [381-383], as well as by an extended use of GWAS on human populations [250-252, 264, 384-389].

The level of pigmentation in the iris, hair and dermis is determined by the amount, type and distribution of melanin in specialized cells, the melanocytes. Melanin is a bio-polymer, produced by melanocytes from tyrosine and catalysed by tyrosinase (TYR). There are two main types of melanin: eumelanin, which is responsible for darker colours, such as black and brown hair, brown eyes and darker skin; and pheomelanin, which determines lighter colours, such as blond and red hair, blue eyes and fair skin [72, 390].

Pigmentation is a polygenic trait regulated by many genes. However, a relatively limited number of genes have a major effect on these traits. Some of these genes, including TYR, TYRP1, DCT, OCA2, MC1R, ASIP and KITLG, are known to be directly involved in the regulation of melanogenesis, while others, such as HERC2, SLC24A5, SLC45A2, SLC24A4, IRF4 and TPCN2 have been associated with pigmentation variation, although their exact interaction and function remain unclear. Table 36 summarizes both the genes and SNPs most significantly associated with human pigmentation traits [14, 72, 384, 387, 390].

Table 36. SNPs and genes significantly associated with human pigmentation traits, listed in the order of prediction significance.

Gene	SNP	Trait associated
HERC2	rs12913832	eye, hair and skin colour
HERC2	rs1129038, rs1667394, rs7183877	eye colour
OCA2	rs1800407	eye and hair colour
OCA2	rs1545397	skin colour
SLC24A4	rs12896399	eye colour

SLC24A4	rs2402130	eye and hair colour
SLC24A5	rs1426654	skin colour
SCL45A2	rs16891982	eye and skin colour
SCL45A3	rs28777	hair colour
SCL45A4	rs16891982	hair and eye colour
TYR	rs1393350	eye colour
TYR	rs1042602	hair colour
TYRP1	rs683	hair colour
IRF4	rs12203592	eye, hair and skin colour
IRF4	rs4959270	hair colour
MC1R	rs885479, rs11547464, rs1805007, rs1805008, rs1805009	skin and hair colour
KITLG	rs12821256	hair colour
ASIP	rs2378249	hair colour
ASIP	rs6119471	skin colour

A number of genetic assays able to predict pigmentation traits have recently been developed [10, 252, 296, 297]. Most of these however, were focused on the European population, providing only a partial spectrum of loci affecting pigmentation away from other population groups. Several recent studies have screened additional populations, confirming known SNPs as well as identifying additional novel markers associated with pigmentation traits [391, 392]. Despite the identification of a significant number of genes and loci affecting the eye, skin and hair colour, the exact interaction between these genes remains unclear [393, 394]. Particularly for the eye and skin colour, it is more likely that a small number of genes with relatively moderate effect such as OCA2, HERC2, SLC24A5 and TYR act together with many genes of a small effect (only partly identified yet), cumulatively explaining the major percentage of heritability [393]. This hypothesis supports the idea that pigmentation prediction should be based on a dense genotype rather than on a small panel of SNPs.

Nevertheless, recent studies demonstrated that a panel of only six SNPs in six pigmentation genes can predict iris colour with particular specificity of the brown and blue shades [292, 293, 297]. In addition, Walsh et. al. demonstrated that 24 SNPs could efficiently predict hair colour [296]. However, some studies argue that the actual prediction accuracy of these assays is significantly lower than claimed due to the high percentage of inconclusive results [394, 395]. The inconclusive results originated mainly in uncertainty of the intermediate eye colour prediction, such as green and hazel.

Natural variation in human skin colour is believed to be a result of positive natural selection as an adaptation to various levels of solar radiation in different habitats [250,

251, 390, 396]. A recent study illustrated that a lighter skin colour in Europeans appeared approximately 11,000 to 19,000 years ago as a mutation from a darker (ancestral) skin colour [397]. Interestingly, two studies in 2014 suggest that a lighter skin colour and hair texture of the evolutionary modern *H. sapiens* (specifically a number of polymorphisms in BNC2 and POU2F3 genes) may have been inherited from Neanderthals through interbreeding and were maintained through evolution, as an adaptation to non-African environments, following exodus from Africa [398, 399]. Notably, a number of polymorphisms in the POU2F3 gene were significantly associated with pigmentation levels as well as with European and Asian ancestries in the current study (detailed in Sections 5.3.4 and 5.3.5).

Despite the fact that pigmentation traits were shown to be under positive evolutionary selection and may strongly correlate to ancestry, estimating ancestry based solely on skin colour is not accurate [15, 250, 251, 400]. There is a strong existing correlation between geographical latitude and skin pigmentation, but the degree of correlation between skin colour and ancestry can vary greatly [371]. Dark skin pigmentation, for example, does not automatically mean that the person is of African ancestry, as many native non-African populations may have dark skin. Conversely, some Asians may have skin as light as many Europeans. Similar skin pigmentation may thus be regulated by different population-specific markers. In order to obtain a complex picture, it is therefore important to screen different populations for pigmentation–associated markers, which was an important aspect of this project.

5.1.3. Craniofacial traits

The human face is the most noticeable of all visible physical traits, and has an extraordinary role in social interactions, medical diagnostics and forensic investigation. The genetic basis of craniofacial morphology has been explored in numerous animal studies with multiple genes shown to be involved as reviewed in Chapter 1 of this document and elsewhere [62, 63, 65, 89, 98, 165, 169-171, 185, 191, 195, 196, 200, 210, 401-409]. Most studies however, have focused on the genetics of craniofacial malformations, rather than providing information on the normal phenotype. The most common disorder associated with various craniofacial abnormalities is craniosynostosis, which symptoms are observed in more than 250 hereditary syndromes [80, 81]. Given that craniosynostosis affects the cranial sutures, genes mutated in this syndrome may

also be involved in influencing normal craniofacial variation and specifically the cranial width and length (reviewed in Section 1.4.2).

A few studies have explored this link and identified several markers significantly associated with cephalic index within the FGFR1 gene [74, 410, 411]. Other studies have focused on cleft lip/palate disorder, affecting the normal facial morphology [144, 412, 413]. A recent study investigated the hypothesis that the same genes that affect non-syndromic clefts may also be responsible for normal phenotype (specifically the bizygomatic distance) and found two SNPs being significantly associated with normal facial variation [160]. Additional studies which focused on the GWAS approach, detected a number of specific polymorphisms in genes that may contribute to normal variation of the nasal area morphology [76, 77]. The limited number of studies focusing on the genetics of normal human craniofacial appearance, illustrates the demand for additional efforts in disclosing the genes and specific markers influencing normal facial variation. This study represents a candidate genes approach, which has not been attempted previously (as reviewed in Chapter 1).

Collecting a large set of craniofacial measurements is essential for performing an association study with candidate genomic markers. This task is tedious, as the anthropometric craniofacial measurements are significantly complex and diverse. Although traditional direct craniofacial measurement methods are considered the anthropometrical "gold standard", they are time consuming, hence less appropriate for a large sample set. Conversely, 3D high resolution technologies, such as laser scanners provide fast and accurate methods for capturing facial landmarks and subsequently calculating a variety of relevant distances using a specialized computer software [41, 76].

Finding specific genetic polymorphisms, affecting facial morphology variation will extend the current limited knowledge on the craniofacial embryogenetics and will enable incorporation of this novel information into a future comprehensive phenotyping assay for forensic intelligence purposes.

5.2. Materials and Methods

5.2.1. AIMs selection

Various online databases and literature sources were screened for potential ancestry informative markers (AIMs) [12, 249, 265, 267, 268, 291, 374, 376, 378, 414-417]. The initial output of these searches represented more than 5,000 AIMs in 1,088 unique genes. These genes were further screened for candidate genes that might be involved in craniofacial development, using web-based bioinformatics tools, as detailed in Material and Methods Chapter 2. The final list included 263 candidate ancestry-informative markers, although due to overlap between AIMs, pigmentation and craniofacial markers, this number is in fact greater, as discussed in Section 5.1.1. Donor ethnicities were recorded based on the self-reported information provided by participants.

5.2.2. Pigmentation markers selection and traits assessment

Genetic polymorphisms, affecting eye, skin and hair colour were selected from previously published sources, as detailed in Chapters 1 and 2.

The pigmentation traits were assigned by a single examiner (the author) according to previously published colour charts. Specifically, eye colour was assessed according to Martin–Schultz scale, hair colour according to Fischer-Saller Scale and finally skin colour according to Fitzpatrick Scale [418-420].

5.2.3. Craniofacial markers and traits selection

Approximately 1,900 genetic polymorphisms were selected as primary genotyping targets within approximately 170 candidate genes and genomic regions, as detailed in Chapter 1. The final amplicon list however, included additional markers found in linkage disequilibrium (LD) with the original SNPs.

A total of ninety two (92) anthropometric measurements, characterizing various facial features and represented by linear and angular distances between craniofacial landmarks, as well as ratios between them, were generated using several computer programs, as discussed in Chapter 2. Additional information about visible traits, including eye lid (single or double); ear lobe (attached or detached); hair texture (straight, wavy, curly or very curly); freckling (none, light, medium or extensive); moles (none, few or many); height, weight and BMI was collected.

5.2.4. Genotyping

Five hundred and eighty seven (587) DNA samples were genotyped for approximately 6,500 genomic markers, using the Ion Torrent sequencing platform, as discussed in Chapter 3. Three samples failed the genotyping step, due to poor sample quantity and/or quality. The sequences were aligned automatically by the Ion Torrent suite software and analysed using the Ion Torrent Reporter software, as discussed in Sections 2.10.4 and 2.10.5.

5.2.5. Statistical analysis

Ancestry inferences were performed using the Structure v2.3.4 software with default parameters as per software developer recommendations [421]. Ancestry estimation analyses were initially performed assuming between two to ten presumed output population clusters (without predefined clusters) or between five to eight pre-defined population clusters. The final analyses were performed assuming four predefined clusters: Europeans, East Asians, Africans and Indians. Analyses were conducted on a reduced set of 523 samples and 1,757 loci following the removal of samples with more than 85% missing alleles. The final ancestry origin was assigned according to "Structure" output. The ancestry origin was estimated as a single (unmixed) source where the main ancestry group could be affiliated with at least 80% of the total mixed ancestry.

Association analyses were performed using SNP & Variation Suite v7 (Golden Helix, Inc., Bozeman, MT) and replicated using PLINK v1.07 software [344]. Statistical

analyses in both software programs implicated linear regression under the assumption of an additive genetic model, while each genotype was binary encoded as 0, 1 or 2. Population stratification correction, incorporated by the EIGENSTRAT function was implemented in SNP & Variation Suite analyses [422, 423]. The PLINK analysis involved usage of a Cochran-Mantel-Haenszel test for the same purpose [424]. In order to reduce any potential confounding effects, all external trait association analyses were performed using sex and BMI compensation algorithm, although for the pigmentation traits, only sex was used as a covariate.

5.2.6. Genes and SNPs annotation analysis

The GeneCards, ENTREZ and UniProtKB/Swiss-Prot web portals were used for annotation analysis of significantly associated genes. The MalaCards web site was used to find potential association between the genes and hereditary syndromes. The GeneMania web site was used to identify a functional network among the genes and encoded proteins. The AmiGo web site was used to find orthologes of unknown Human genes in other organisms. Mouse genome database (MGI) was used to search for genes phenotype in relevant craniofacial mouse mutants SNPnexus and Alfred websites were used to annotate SNPs.

The SNP Annotation and Proxy Search (SNAP) web portal was used to find SNPs in linkage disequilibrium (LD) and generate LD plots, based on the CEU population panel from the 1000 genomes data set, within a distance of up to 500kb and r^2 threshold of 0.8 [425].

The Regulome database and potentially functional database (PFS) searches were implemented to annotate SNPs with known and predicted regulatory elements in the intergenic regions of the *H. sapiens* genome [333, 426].

5.3. Results and Discussion

The following section aims to summarise the genetic association studies of ancestry, pigmentation and craniofacial traits. Following these association results, predictive statistical analyses undertaken for each of the specific traits are presented (Section 5.4).

5.3.1. Sample descriptive statistics

The following section briefly outlines the numbers and their percentage proportions of samples used for each trait in the present study.

5.3.1.1. Ancestry origin descriptive statistics

The majority of the sample set was composed of 363 self-declared Caucasians. These samples were merged with additional 29 samples from the Middle Eastern population to satisfy analysis purposes, as there was a relatively limited sample number of the latter. Additional population samples included 56 East Asians, 45 South Asians (Indians), 23 Africans, 3 Aboriginals and 4 samples of other population groups. Sixty one (61) samples were defined as "Admixture", based on ancestry information provided by volunteers (Table 37).

Table 37. Sample numbers as categorised by self-reported ancestry.

Ancestry	Number of samples	Percentage
Caucasian	363	62.16
E. Asian	56	9.59
African	23	3.94
Aboriginal	3	0.51
Indian	45	7.71
Admixture	61	10.45
Other	4	0.68
Middle Eastern	29	4.97
Total	584	100

The majority of genetic association studies are usually performed on relatively homogenic (population-wise) sample sets [427]. The reason for this is that association analysis performed on a diverse sample set may introduce substantial bias due to etiological population stratification and may produce spurious (false positive) associations. Population stratification results from the fact that allele frequencies in different populations may vary significantly due to their respective biological adaptation. These are driven by positive selection throughout the evolution process, increasing population differentiation at specific genomic regions [251, 253, 254, 257, 262, 428]. As a result, risk alleles or quantitative trait loci (QTL) in this specific case, may differ in their occurrence across populations or even be absent in certain ethnic groups. This plays a critical role in assessing risk factors of certain medical conditions in a developing personalized medicine field as well as in accurate prediction of EVTs in the forensic context.

The relative diversity of samples used in this study permitted the discovery of genetic associations and subsequent prediction of phenotypic traits in various population groups. Therefore, performing both accurate ancestry estimation and statistical correction for population stratification was an important step in this study (as detailed in Sections 5.2.5 and 5.3.2)

5.3.1.2. Pigmentation traits descriptive statistics

Eye pigmentation was grouped into four main colour categories that included brown, blue, green and hazel (Table 38). Not surprisingly, brown eyed individuals accounted for half of the dataset, since brown is dominant and the most common iris colour in the world [429]. The next major group was represented by individuals with blue eyes (28.5%) and the rest by intermediate colours such as green and hazel.

Approximately half of the participants had brown hair (Table 38), almost 28% had black hair, 17.5% had blond hair, while red haired individuals (including various shades of red such as auburn, carrot, orange and strawberry) contributed only almost 4% of the sample set.

The fair and average skin colours, which are often hard to distinguish, represented together approximately 67% of the samples. The dark skin colours (black and brown)

contributed almost 20%, and the intermediate "olive" colour approximately 13% of the total sample numbers.

Three additional shades of each colour (light, medium and dark) were used to describe the pigmentation pattern. However, these sub-categories were not used for the purpose of the association and prediction studies, as this study has focused only on the main pigmentation shades.

Table 38. Eye, skin and hair colour distribution among genotyped samples.

Eyes colour	Number of samples	Percentage
Brown	291	50.0
Blue	166	28.5
Green	65	11.2
Hazel	60	10.3
Total	582	
Hair colour	**Number of samples**	**Percentage**
Black	163	27.9
Brown	296	50.7
Blonde	102	17.5
Red	23	3.9
Total	584	
Skin colour	**Number of samples**	**Percentage**
Fair	173	29.7
Average	216	37.1
Olive	78	13.4
Black	14	2.4
Brown	101	17.4
Total	582	

Contrary to many published studies where skin, hair and eye colours were self-reported or assessed by multiple examiners, pigmentation phenotype assessment in this project was performed by a single examiner (the author). This approach should decrease any bias introduced by subjective phenotype assessment of participants. Nonetheless, a "single examiner" approach lacks objective standardization as it is based on the examiner's ability to clearly distinguish pigmentation patterns. This problem can

potentially be solved by using digital methods (such as spectrophotometry) for colour quantification. A Chroma Meters CR-400 by Konica Minolta was tested in this study for this purpose, but due to possible eye damage, it was not eventually pursued in this current project.

5.3.1.3. Craniofacial traits descriptive statistics

A comprehensive summary of the craniofacial trait – related statistics is detailed in Section 4.4.

5.3.2. Ancestry estimation using STRUCTURE

A significant difference in allele frequencies among populations (Fst) emphasizes an importance of accurate ancestry estimation, prior to undertaking any trait association analyses, as discussed in Section 5.3.1.1 and reviewed in other studies [430]. Self-reported ancestry however, cannot be considered a fully reliable source due to either lack of accurate information or misleading information on individual's report on family history [431, 432]. Additionally, some of the population clusters commonly used in genetic association studies (e.g. Caucasians or Africans) are not always represented by genetically homogeneous groups and may display significant inter-population diversity. For example, a recent study in 2009 investigated 121 African populations and 60 non-African populations and found a higher genetic diversity between African populations compared to non-African populations, among other differences [433]. Self-declared ancestry should therefore be confirmed by analysis of the ancestry-informative genetic markers, which would aim on providing a more reliable and less biased ancestry estimation. The most widely used software for this purpose is the Structure [421]. Population analysis using the Structure package enables to visualize the ancestry composition of samples, indicating potential admixture, and categorize them more accurately into specific clusters, allowing reliable analysis of potential genomic associations. This software was widely used in numerous studies, including forensically – focused publications [246, 265, 434, 435]. The Structure clustering model assumes "k" populations, each of which is characterized by a set of allele frequencies at each locus. Analysed samples are assigned (probability-wise) to specific population clusters,

or jointly to two or more populations if there is an indication of admixture. Structure algorithms can be applied to most of the commonly used genetic markers, if they are not closely linked.

Ancestry estimation was performed using all 3,477 genetic markers (not only initial AIMs), following sequencing quality control filtering (DP and GQ >10) and MAF≥2%, as discussed in Section 5.3.3. The initial Structure analysis was performed with various numbers of predefined population clusters (up to nine), incorporating admixture model and prior population information. This clustering made it possible to test how well the software can estimate (predict) the ancestry origin. This test revealed that four and five clusters (k=4 or k=5) produced the most informative data output. The five output clusters grouped samples as follows with (1) Caucasian, (2) East Asian, (3) African, (4) South Asian (Indian) and (5) Aboriginal. The program assigned resulting samples to either the original (identical to self-reported) ancestry clusters (based on up to 20% admixture threshold), or to the Admixture group, if major ancestry contribution was below 20 percent. Ancestry estimation based on the five pre-defined clusters resulted in the following number of samples: Caucasian (n=365), Asian (n=50), African (n=16), Indian (n=40), Aboriginal (n=7) and the rest estimated as an admixture of these population groups (Figure 55).

Notably, a significant proportion of Caucasian samples showed various levels of admixture with Aboriginal and Indian population groups. This finding may reflect recent Australian history, although a more stringent analysis with a larger sample size, especially of the Indigenous samples, is needed to test this hypothesis.

Figure 55. Structure output visualized as a color-coded Q plot, based on five pre-defined population clusters. X axis – overall percentage, Y axis – clusters (k). The colours represent: red – Caucasian, green – Asian, Blue – African, Yellow –Aboriginal and Purple – Indian ancestry clusters.

Due to a limited number of Aboriginal samples, the final Structure analysis was performed assuming only four population groups, excluding the Aboriginal cluster. After removing Aboriginal ancestry samples and samples that were missing more than 80% of genotyped markers, a total number of 516 samples remained. The majority of the remaining samples were distinctly categorized in clear clusters, in general concordance with their original (self-defined) ancestry. However, several samples were assigned to a new (different to the self-reported) ancestry cluster (Figures 56 and 57). In summary: 459 samples (89%) tested with Structure were assigned the same ancestry cluster (sole or mixed origin) as initially collected self-reported ancestry information. Of the remaining 57 individuals, a total number of 39 individuals were estimated as an 'Admixture' (based on up to 20% admixture threshold) and 18 samples were assigned a different single ancestry, when compared to the original self-reported ancestry (Table 39). Of the 39 newly estimated 'Admixture' samples, 16 samples were originally self-reported as Middle Eastern, 12 samples as Caucasian, seven (7) samples as Indian and four (4) as African. This is not surprising for the Middle Eastern and Indian ancestry individuals, as the geographic regions such as Indian peninsula and especially Levant are known for extensive demographic movement and admixture [436]. Interestingly, two samples of Russian descent who were self-declared as Caucasian, were estimated as Admixture by the Structure. This outcome can be explained by the population admixture, that possibly historically resulted from Mongol invasion that lasted from 1237 AD to 1480 AD.

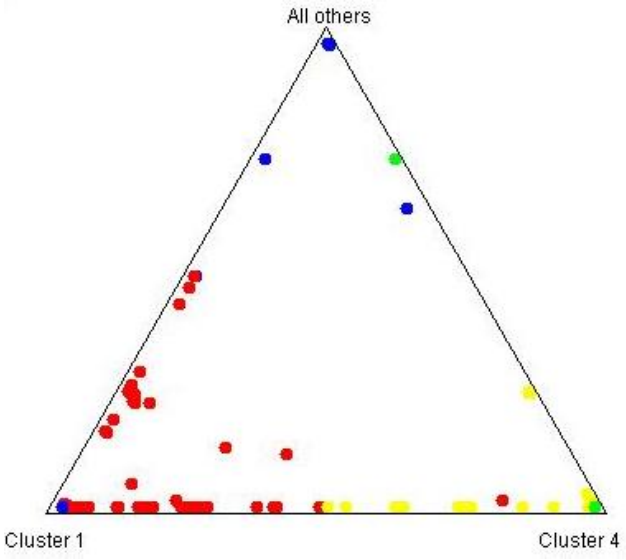


Figure 56. Triangle plot of the sample subset tested with STRUCTURE. The colours represent: red – Caucasian, green – Asian, Blue –African and Yellow – Indian ancestry clusters. Some samples show same clustering within their self-reported ancestry clusters, while others demonstrate different clustering pattern.

Figure 57. STRUCTURE output visualized as a color-coded Q plot, based on four pre-defined population clusters. X axis – overall percentage, Y axis – clusters (k). The colours represent: red – Caucasian, green – Asian, Blue – African and Yellow – Indian ancestry clusters. Note that the yellow colour in this figure represents a different cluster comparing to Figure 55.

Following Structure data output, of the 18 samples predicted differently (non-admixed ancestry), ten (10) were from the original (self-reported) Middle Eastern sub-population. These samples were re-clustered as European, based on Structure prediction. The majority of these individuals were either of Lebanese or Greek background. Interestingly, all the Lebanese individuals who were estimated as Europeans by Structure were self-declared Arab Christians. This result may illustrate the evidence of an European population admixture with the local populations during the Crusader period in this region [437].

In summary, the self-reported Admixture cluster was mostly affected by Structure algorithm prediction. This group has grown from 61 samples (based on the self-reported ancestry) to 107 samples (based on the Structure prediction). Following the removal of a mixed ancestry samples from the Caucasian cluster, it was renamed 'European', based on its homologous population content.

New ancestry clusters estimated by Structure were incorporated into the SVS GoldenHelix and PLINK softwares and used for association analyses, as detailed in Sections 5.3.3 and 5.3.4. It should be noted however, that Structure prediction is mainly based on the differences in allele frequencies between populations (Fst) and its "resolution" is largely dependent on the differentiation ability of the specific markers used for ancestry estimation. As a result, this approach is not error-free and must be treated with caution. Nevertheless, ancestry prediction by Structure is considered as less biased and more accurate than self-reported ancestry [430, 434]. In this study, ancestry prediction for the majority of samples was concordant with the self-reported source. Given that many people are not fully aware of their family history, re-clustering of their ancestry by Structure to the Admixture cluster was considered beneficial for this study.

Only 18 samples that were clustered into a different (single) ancestry group were in a conflict with their self-declared ancestry. In these cases, Structure prediction was considered more reliable. Table 39 provides a summary of sample numbers based on Structure-estimated ancestries and on the original self-reported ancestries.

Table 39. Final ancestry statistics, estimated by Structure software and used for the association analysis.

Self-reported ancestry	Number of samples	Percentage
Caucasian	363	62.16
E. Asian	56	9.59
African	23	3.94
Aboriginal	3	0.51
Indian	45	7.71
Admixture	61	10.45
Other	4	0.68
Middle Eastern	29	4.97
Total	584	100

Structure- estimated ancestry	Number of samples	Percentage
European	367	62.84
E. Asian	51	8.73
African	16	2.74
Indian	43	7.36
Admixture	107	18.32
Total	584	100

5.3.3. Genotyping statistics and initial filtering of the data

Five hundred and eighty seven (587) DNA samples were genotyped by sequencing using the Ion Torrent platform, as detailed in Chapter 2. Three (3) samples did not produce results due to either poor DNA quality or unsuccessful library and template preparation. Targeted genotyping of 584 samples resulted in 9,051 unique markers, although the majority of markers (>5,000) were represented by rare polymorphisms of 1% MAF or less (Figure 58).

An essential step of data quality control was performed by removing markers of low genotype quality (GQ>10) and sequencing depth (DP>10), which ultimately resulted in 8,229 markers. The majority of the filtered markers were represented by relatively rare minor frequency alleles (MAF<5%). Figure 58 shows the distribution of minor allele frequency and call rates among samples prior to MAF pruning.

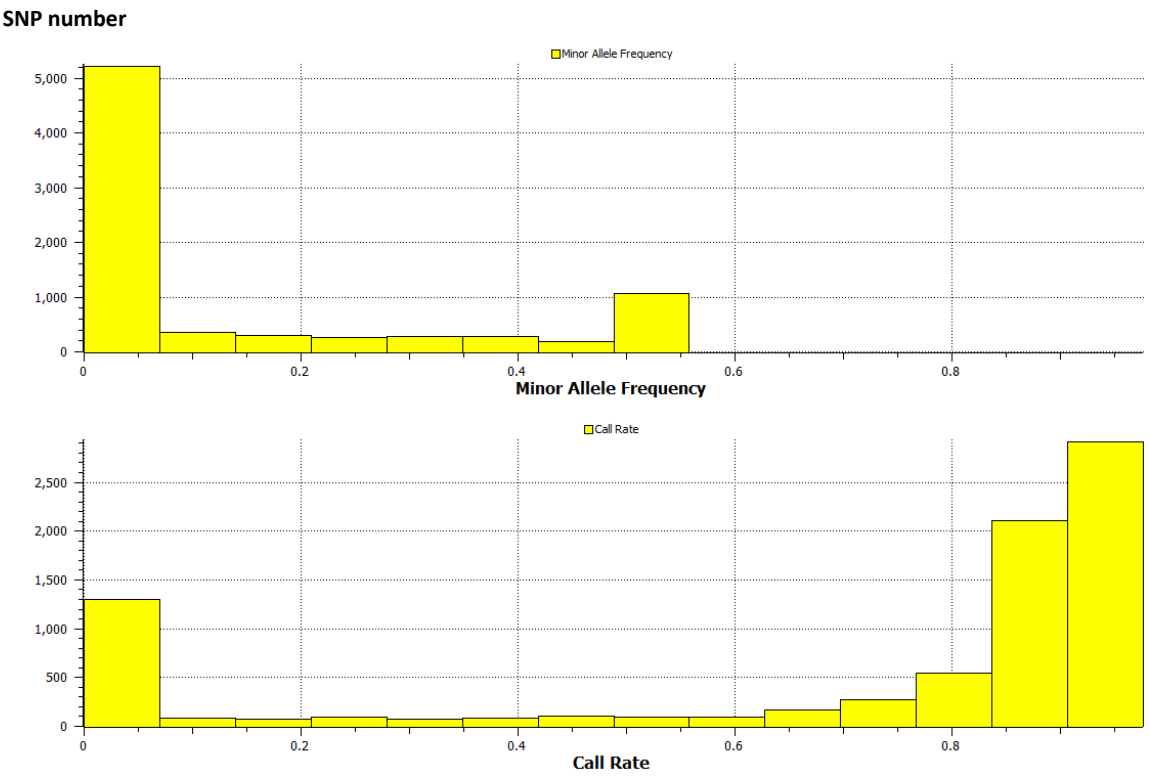


Figure 58. Allele frequencies and call rate distribution among genotyped samples, prior to filtering (n=9,051).

Filtering of markers by various levels of MAF (1%, 2% and 3%), produced 3,653, 3,477 and 3,367 markers respectively, while a 2% MAF cut-off was considered the most appropriate, based on the literature [438, 439]. Additional filtering based on the Hardy-Weinberg equilibrium (HWE) threshold of 1E-07, resulted in 2,858 markers.

An important criterion for assessing the association results is the accuracy of genotyping, which can be measured by the sequencing depth. The average sequencing depth for the significantly associated markers in this study was approximately x57. This is a high sequencing depth, particularly compared to only x4, which was orriginally used for the first published stages of the Hapmap and the 1000 genomes genotyping projects, or x30 depth, used for their subsequent stages [245, 440].

5.3.4. AIMs association analysis

Association analyses performed on four (4) main population groups (European, East Asian, African and Indian) produced similar results in both SVS and PLINK software. However, the relative ranking of several statistically significant SNPs and their relative

p-values was the only slight difference that could be noted. The results presented in Tables 40-43 are based on the SVS analyses. Given the very large number of significant markers that were detected, only the top twenty SNPs are listed.

Table 40. Twenty top SNPs and respective genes found to be associated with European ancestry. Chromosomal location, average sequencing depth per SNP, rs number of each marker and original and Bonferroni –corrected p-values are included. SNPs found in linkage disequilibrium are highlighted in yellow.

Marker	Average Depth	rsID	Gene	Full-Model P-Value	Bonferroni P
5:33951693	51.9114	rs16891982	SLC45A2	1.01E-66	2.07E-63
15:28365618	63.5959	rs12913832	HERC2	2.34E-27	4.79E-24
17:19175317	65.7768	rs28591622	EPN2	1.83E-26	3.76E-23
17:19239432	42.9114	rs1043809	EPN2	3.22E-24	6.60E-21
17:19204863	56.3708	rs4924980	EPN2	1.80E-23	3.69E-20
2:152829657	91.5424	rs6721518	CACNB4	2.48E-23	5.09E-20
17:19172505	37.4594	rs28760541	EPN2	2.23E-21	4.58E-18
15:48426484	47.4004	rs1426654	SLC24A5	1.35E-20	2.77E-17
17:19211073	40.4834	rs8072587	EPN2	4.25E-20	8.72E-17
11:120133494	98.3911	rs2715883	POU2F3	1.79E-19	3.67E-16
12:66168151	69.1679	rs7134682	near RPSAP52	1.81E-19	3.72E-16
5:33969628	48.3911	rs35414	SLC45A2	2.56E-19	5.25E-16
17:19224397	62.5221	rs6587216	EPN2	3.78E-19	7.75E-16
13:102816760	59.7048	rs2607642	FGF14	5.31E-19	1.09E-15
11:120130512	46.8339	rs1941411	POU2F3	5.04E-18	1.03E-14
17:19247075	43.8229	rs4924987	B9D1	1.02E-17	2.09E-14
5:33958910	65.0609	rs1010872	SLC45A2	1.60E-17	3.28E-14
13:102821327	44.81	rs4771420	FGF14	1.70E-17	3.48E-14
2:152829626	92.0793	rs16830527	CACNB4	1.76E-17	3.62E-14
5:33954511	80.2694	rs2287949	SLC45A2	2.09E-17	4.29E-14

Table 41. Twenty top SNPs and respective genes found to be associated with Asian ancestry. Chromosomal location, average sequencing depth per SNP, rs number of each marker and original and Bonferroni –corrected p-values are included. SNPs found in linkage disequilibrium are highlighted in yellow.

Marker	Average Depth	rsID	Gene	Full-Model P-Value	Bonferroni P
22:46835000	99.9133	rs4823810	CELSR1	5.98E-79	1.23E-75
1:36437991	73.6494	rs12096413	EIF2C3	8.36E-70	1.71E-66
1:36426398	28.9502	rs7521611	EIF2C3	2.71E-60	5.56E-57
1:36099200	80.6328	rs6668101	PSMB2	1.31E-53	2.69E-50
11:120129533	55.4446	rs11217777	POU2F3	9.65E-52	1.98E-48
11:120169962	41.0203	rs11217806	POU2F3	1.23E-51	2.52E-48
11:120121323	45.8118	rs11217775	POU2F3	5.38E-51	1.10E-47
2:109556761	63.1882	rs260710	EDAR	2.60E-50	5.34E-47
12:112843363	49.4852	rs2301723	RPL6	7.25E-50	1.49E-46
11:120170030	43.5	rs11217807	POU2F3	1.66E-49	3.40E-46
15:48426484	47.4004	rs1426654	SLC24A5	1.77E-49	3.62E-46
1:36424338	43.6421	rs716924	EIF2C3	1.92E-48	3.93E-45
8:73848139	63.9317	rs3735829	KCNB2	9.89E-46	2.03E-42
1:234330571	58.7768	rs7531988	SLC35F3	9.00E-45	1.84E-41
2:109556667	66.5683	rs260709	EDAR	1.65E-44	3.37E-41
11:120122394	39.4336	rs11827471	POU2F3	3.47E-43	7.12E-40
2:109579738	62.7011	rs260690	EDAR	6.86E-43	1.41E-39
13:34864240	64.5738	rs2065982	between RFC3 and GAMTP2	8.55E-43	1.75E-39
2:109616376	35.0627	rs17034770	near EDAR	1.52E-42	3.11E-39
2:109586371	63.5959	rs11123719	EDAR	3.45E-41	7.06E-38

Table 42. Twenty top SNPs and respective genes found to be associated with Indian ancestry. Chromosomal location, average sequencing depth per SNP, rs number of each marker and original and Bonferroni –corrected p-values are included.

Marker	Average Depth	rsID	Gene	Full-Model P-Value	Bonferroni P
5:33951693	51.9114	rs16891982	SLC45A2	3.17E-17	6.49E-14
5:33958910	65.0609	rs1010872	SLC45A2	4.99E-16	1.02E-12
5:33954511	80.2694	rs2287949	SLC45A2	3.57E-14	7.31E-11
17:19157989	88.0554	rs4924960	EPN2	5.48E-10	1.12E-06
17:19163690	35.8007	rs4924778	EPN2	3.38E-09	6.93E-06
11:113291752	80.5683	rs117431082	DRD2	5.98E-09	1.23E-05
2:152814028	45.7177	rs16830498	CACNB4	1.05E-08	2.16E-05
2:152829626	92.0793	rs16830527	CACNB4	1.57E-08	3.22E-05
15:28365618	63.5959	rs12913832	HERC2	1.82E-08	3.74E-05
12:66256395	58.5129	rs11175944	HMGA2	5.41E-08	1.11E-04

9:12672320	36.5609	rs1408801	near TYRP1	5.56E-08	1.14E-04
5:33969628	48.3911	rs35414	SLC45A2	1.20E-07	2.45E-04
13:111827167	87.3155	rs9522149	ARHGEF7	1.94E-07	3.97E-04
5:61013725	87.0923	rs16894149	-	2.14E-07	4.39E-04
1:27931698	49	rs4908343	AHDC1	2.24E-07	4.59E-04
9:12672396	36.6937	rs9919017	-	2.67E-07	5.47E-04
12:66260924	62.131	rs343087	HMGA2	3.89E-07	7.98E-04
12:66236735	27.1974	rs2272047	HMGA2	5.12E-07	1.05E-03
17:19172505	37.4594	rs28760541	EPN2	1.05E-06	2.15E-03
12:66256395	58.5129	rs11175944	HMGA2	5.41E-08	1.11E-04

Table 43. Twenty top SNPs and respective genes found to be associated with African ancestry. Chromosomal location, average sequencing depth per SNP, rs number of each marker and original and Bonferroni –corrected p-values are included.

Marker	Average Depth	rsID	Gene	Full-Model P-Value	Bonferroni P
7:83030169	26.3506	rs2709927	SEMA3E	3.53E-118	7.23E-115
18:67672106	58.8616	rs56293475	RTTN	1.14E-116	2.34E-113
10:34755348	32.3266	rs1978806	PARD3	7.91E-109	1.62E-105
2:232198375	73.4834	rs6747710	ARMC9	2.71E-104	5.56E-101
7:41986689	82.1162	rs7789366	near GLI3	1.32E-98	2.71E-95
8:72131359	68.3893	rs73684719	EYA1	1.09E-88	2.24E-85
7:107704688	66.2269	rs10257477	LAMB4	3.06E-87	6.26E-84
7:83033303	34.321	rs2722980	SEMA3E	8.38E-83	1.72E-79
7:83041344	60.8137	rs2709963	SEMA3E	1.41E-80	2.88E-77
7:107696289	64.7565	rs17154865	LAMB4	7.21E-78	1.48E-74
8:72127562	34.5258	rs79867447	EYA1	1.42E-77	2.91E-74
8:116700847	40.9004	rs7842702	TRPS1	3.56E-77	7.30E-74
8:116776330	42.4465	rs10505268	TRPS1	5.44E-76	1.11E-72
20:45801340	61.5517	rs6090632	EYA2	3.36E-70	6.88E-67
18:19765739	45.5203	rs16964572	GATA6	5.15E-70	1.06E-66
7:83041372	59.6753	rs2709964	SEMA3E	2.11E-69	4.32E-66
10:24195798	26.6513	rs73604433	KIAA1217	2.99E-69	6.12E-66
2:177992107	52.7546	rs6433650	near RPL29P8	1.31E-68	2.69E-65
6:137345768	34.5203	rs276488	IL20RA	4.16E-67	8.54E-64
2:232198327	74.1716	rs6719593	ARMC9	6.48E-67	1.33E-63

Based on the Bonferroni corrected p-value threshold of <0.05, 215 markers were significantly associated with the European ancestry, 495 with the Asian ancestry, 627 with the African ancestry and 41 with the Indian ancestry. The relatively small number of SNPs associated with Indian population could be explained by the fact that AIMs used in this study were selected from a number of publications that focused primarily on the Caucasian, Asian and African populations (as detailed in Section 5.1.1).

Given that most GWAS studies use a more stringent threshold of between 1E-05 to 1E-08 of the non-corrected p-value, a higher threshold of unadjusted p-value was considered. An application of a threshold of 1E-07 resulted in reduction of statistically significant markers from 215 to 142 for the European, from 495 to 364 for the Asian, from 627 to 528 for the African and from 41 to 19 for the Indian.

Based on the top 20 significant markers for each ancestry, 72 non-tagged significant SNPs in 35 genes and genetic regions were detected (Tables 40-43). The strongest (Bonferroni-corrected) significance of association was demonstrated for the African ancestry cluster (down to 7.23E-115 p-value for the top SNP) and the lowest for the Indian population (down to 6.49E-49 p-value for the top SNP). Given the stringent Bonferroni correction, these results demonstrate strong association between each of the four ancestry groups and specific genetic markers, tested in this study. The most significant similarities in both SNPs and genes were observed between the European and the Indian populations (4 genes and 7 SNPs). In contrary, the markers and genes in the Asian and African populations showed no overlap. The existence of a common Indo-European proto-language and relevant genetic studies suggests that European and Indian populations are indeed genetically more similar, compared to Africans and East Asians [441, 442]. Notably, the major similarity in the craniofacial measurements in this dataset was observed between the Indian and the West European population groups, as discussed in Section 4.5. On the other hand, this observation may be due to population heterozygosity (Fst) of the specific markers used in this study, which were unable to distinguish between these population groups.

The two SNPs most significantly associated with European and three markers most significantly associated with Indian population groups were found in genes previously associated with eye or skin pigmentation as SLC45A2 and HERC2 genes. This is likely the result of an overlap between ancestry and pigmentation-informative markers (as discussed in Section 5.1). For example, rs16891982 in SLC45A2 gene was the top SNP found to be associated with European ancestry (Leu374Phe amino acid change). This allele was previously associated with light-skin European ancestry [443]. Another

polymorphism in this gene was documented to be in association with light skin and with protection from malignant melanoma within European population (F374L amino acid change) as well as with black hair phenotype [386, 444, 445]. The rs12913832 in the HERC2 gene was found as the second mostly significant marker associated with the European ancestry and it was also identified as the most important SNP for eye pigmentation variation (brown versus blue eyes) [446, 447]. While the pigmentation genetics is the primary association of these markers, they are clearly informative for ancestry prediction, although indirectly.

Other phenotypic traits may also provide complementary information on ancestry. For example, Asian ancestry associated gene EDAR, was previously shown to be associated with hair morphology (thick hair) and this trait is especially common in the Asian population [448, 449]. Polymorphisms in several genes such as TRPS1, ARMC9, GLI3, LAMB4 and EYA1 were found significantly associated with the African ancestry in this study, as well as with very curly hair (see Table 45). Interestingly, mutations in the TRPS1 gene were previously identified in the Hypertrichosis (also called Ambras syndrome), a condition that is associated with abnormal amount of hair growth over the body [450]. The association of this gene with very curly hair is a novel finding and has not been described previously. It should be emphasized however, that the observed association could be a result of relatively small African sample set and should be subsequently checked in replication study with larger sample set .

Additional significantly associated genes, such as LAMB4, RTTN, SEMA3E and CELSR1 were previously shown to have a function (known or unknown) in embryonic development regulation, including craniofacial embryogenesis in human or model organisms [323]. Notably, these genes also demonstrated association with craniofacial traits in the current study (as discussed in Section 5.3.6).

The SVS genomic browser analysis showed that almost all of the significantly associated markers are located in introns or in intergenic regions, in concordance with candidate markers selection process and as discussed elsewhere [261]. The Regulome database analysis which predicts regulatory elements in intergenic SNPs, revealed that 43 of 73 top SNPs in all tested ancestries have a potential role in either regulating DNAase hypersensitivity, affecting binding sites of transcription factors and altering promoter regions. A search using potentially functional database (PFS), which predicts regulatory elements in the intragenic and intergenic regions resulted in 68 SNPs that may have potential function in regulation of various cellular processes.

Based on SNAP analysis, numerous markers located in the RTTN, POU2F3 and EDAR genes were found to be in linkage disequilibrium, (highlighted in Tables 39 and 41 and Figure 59) and potentially representing a haploblock. Tag SNP analysis of 215 markers that were significantly associated with the European ancestry resulted in 48 tag SNPs. The use of these tag SNPs provided an almost five-fold reduction of markers, offering greater flexibility for the use of different genotyping platforms with limited capabilities, without decreasing prediction accuracy.

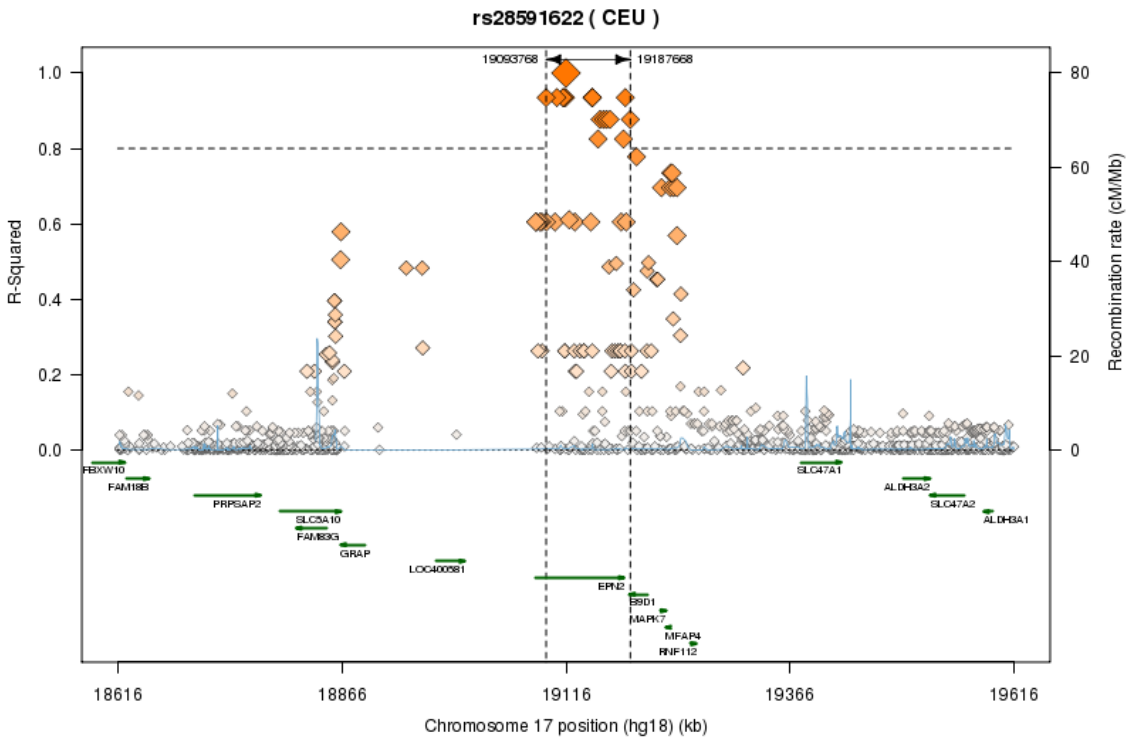


Figure 59. An example of LD plot, visualizing the top 20 SNPs associated with Caucasian ancestry, generated by SNAP.

5.3.5. Pigmentation traits association study

This section summarises the association analysis of the pigmentation traits. The results are organized in three sub-sections, according to eye, skin and hair colour with a separate discussion for each section.

In general, eye, skin and hair association results demonstrated an overlap in markers between these traits, as well as with the ancestry informative markers, as discussed in Section 5.3.4. The "key" pigmentation genes that have a major impact on the pigmentation, such as HERC2, OCA2, SLC45A2, SLC24A5 produced the most significant associations with all pigmentation traits, consistent with published sources.

The average SNP sequencing depth was approximately x57. This high sequencing depth provides additional confidence in the association results of this study.

5.3.5.1. Eye colour

The genetics of the eye pigmentation is better understood relatively to skin and hair. Several genes such as HERC2, OCA2, SLC24A4 and SLC24A5 are known to play a central role in iris pigmentation patterning, while a single SNP rs12913832 in the HERC2-OCA2 intergenic region has the highest prediction ability of blue/brown eye colour [296].

Eye colour in this study was grouped into four main categories that included brown, blue, green and hazel, and analysed for potential association with a set of candidate SNPs. The summary of the eye colour association results are presented in Table 44.

Table 44. A summary of eye colour association study.

Trait	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
Brown eyes	15:28365618	63.5959	rs12913832	HERC2	6.41E-71	1.30E-67
	5:33951693	51.9114	rs16891982	SLC45A2	2.45E-28	4.98E-25
	15:28356859	69.3697	rs1129038	HERC2	1.83E-26	3.72E-23
	15:28516084	53.155	rs8039195	HERC2	2.50E-23	5.08E-20
	15:28344238	64.7122	rs7495174	OCA2	8.34E-19	1.69E-15
	15:28513364	50.3991	rs916977	HERC2	3.04E-15	6.17E-12

	15:48426484	47.4004	rs1426654	SLC24A5	1.25E-14	2.55E-11
	15:28268990	43.7915	rs749846	OCA2	3.46E-10	7.02E-07
	15:28200408	43.5111	rs7170989	OCA2	5.35E-10	1.09E-06
	15:52611451	61.0443	rs2290332	MYO5A	3.45E-09	7.00E-06
	15:28231279	45.5018	rs4778221	OCA2	6.38E-09	1.30E-05
	5:33958910	65.0609	rs1010872	SLC45A2	2.17E-08	4.41E-05
	9:87772807	55.7546	rs526454	between NTRK2 and STK33P1	2.20E-08	4.47E-05
	5:33954511	80.2694	rs2287949	SLC45A2	1.21E-07	2.45E-04
	10:55949899	34.7048	rs16905691	PCDH15	1.37E-07	2.78E-04
	17:19174874	76.6417	rs1467028	EPN2	1.69E-07	3.42E-04
	3:58112440	56.6937	rs2362903	FLNB	2.03E-07	4.13E-04
	2:240282208	86.7675	rs12471054	HDAC4	2.10E-07	4.27E-04
	1:36424338	43.6421	rs716924	EIF2C3	2.52E-07	5.11E-04
	12:112843363	49.4852	rs2301723	RPL6	4.07E-07	8.26E-04
	2:109556761	63.1882	rs260710	EDAR	4.97E-07	1.01E-03
	13:111827167	87.3155	rs9522149	ARHGEF7	5.87E-07	1.19E-03
	11:120133494	98.3911	rs2715883	POU2F3	6.54E-07	1.33E-03
	12:66115201	52.0996	rs12300373	between PCNPP3 and RPSAP52	7.01E-07	1.42E-03
	2:152814028	45.7177	rs16830498	CACNB4	7.02E-07	1.42E-03
	11:120170030	43.5	rs11217807	POU2F3	7.03E-07	1.43E-03
	2:109557099	54.6218	rs260711	EDAR	8.80E-07	1.79E-03
	1:36437991	73.6494	rs12096413	EIF2C3	9.35E-07	1.90E-03
	10:55968685	64.7048	rs10825273	PCDH15	9.88E-07	2.01E-03
Blue eyes	15:28365618	63.5959	rs12913832	HERC2	2.38E-39	4.82E-36
	15:28356859	69.3697	rs1129038	HERC2	8.55E-27	1.74E-23
	15:28513364	50.3991	rs916977	HERC2	1.35E-16	2.75E-13
	15:28516084	53.155	rs8039195	HERC2	2.66E-16	5.39E-13
	15:28365733	108.0498	rs7183877	HERC2	6.41E-09	1.30E-05
	5:33951693	51.9114	rs16891982	SLC45A2	1.10E-08	2.23E-05
	15:28344238	64.7122	rs7495174	OCA2	7.06E-08	1.43E-04
	15:28268990	43.7915	rs749846	OCA2	7.21E-07	1.46E-03
	14:92781001	44.607	rs4904868	between CPSF2 and SLC24A4	1.95E-06	3.96E-03
	14:92773903	39.8911	rs12896471	-	7.32E-06	1.49E-02
	Y:15026424	13.885	rs2032624	DDX3Y	8.76E-06	1.78E-02
	15:28200408	43.5111	rs7170989	OCA2	1.10E-05	2.24E-02
Green eyes	15:28365618	63.5959	rs12913832	HERC2	6.87E-11	1.41E-07
	5:33951693	51.9114	rs16891982	SLC45A2	4.25E-07	8.71E-04
	5:33969628	48.3911	rs35414	SLC45A2	6.04E-06	1.24E-02
Hazel eyes	15:28356859	69.3697	rs1129038	HERC2	2.27E-06	4.65E-03

The most significantly associated markers with eye colour were in general consensus with the published sources. Following the application of 1E-07 Bonferroni un-adjusted p-value threshold (due to high number of associated markers), brown eye colour association analysis produced the majority of significant results, revealing 29 significant markers in 17 genes and genomic regions. It was followed by the blue eye colour, showing association with 12 markers in five genes. Green eye colour was found associated with three markers in two genes, while the hazel eye colour was associated with only one marker. Not surprisingly, both intermediate colours revealed less significant p-values of associated markers compared to the brown and blue colours [394, 395]. The outcome of this genetic association analysis can be clustered into three main groups, according to the functional role of the identified gene (as summarized in Table 45).

The most significant markers were found in genes which have a major impact on pigmentation such as HERC2, OCA2, SLC45A2, SLC24A5 and MYO5A. Among these markers, the lowest p-value was associated with rs12913832 in the HERC2-OCA2 genomic region, followed by rs16891982 in the SLC45A2 and rs1129038 in the HERC2 gene (as illustrated in Figure 60). These markers are known to be the key players in eye pigmentation regulation and most significant predictors of the blue and brown eye colours, as discussed in Section 5.1.2.

Additional markers were found in several genes that were not previously shown to be associated directly with the eye colour. However, these genes were shown to be involved in melanogenesis or be expressed in melanoma, hence potentially affecting pigmentation regulation. Another group of genes can be considered ‘novel’ pigmentation genes, as no potential association of these genes with pigmentation (either eye, skin or hair) has been demonstrated to date. A summary of the mostly significant associated genes is shown in Table 45. The electronic version of this table includes web links to the relevant references.

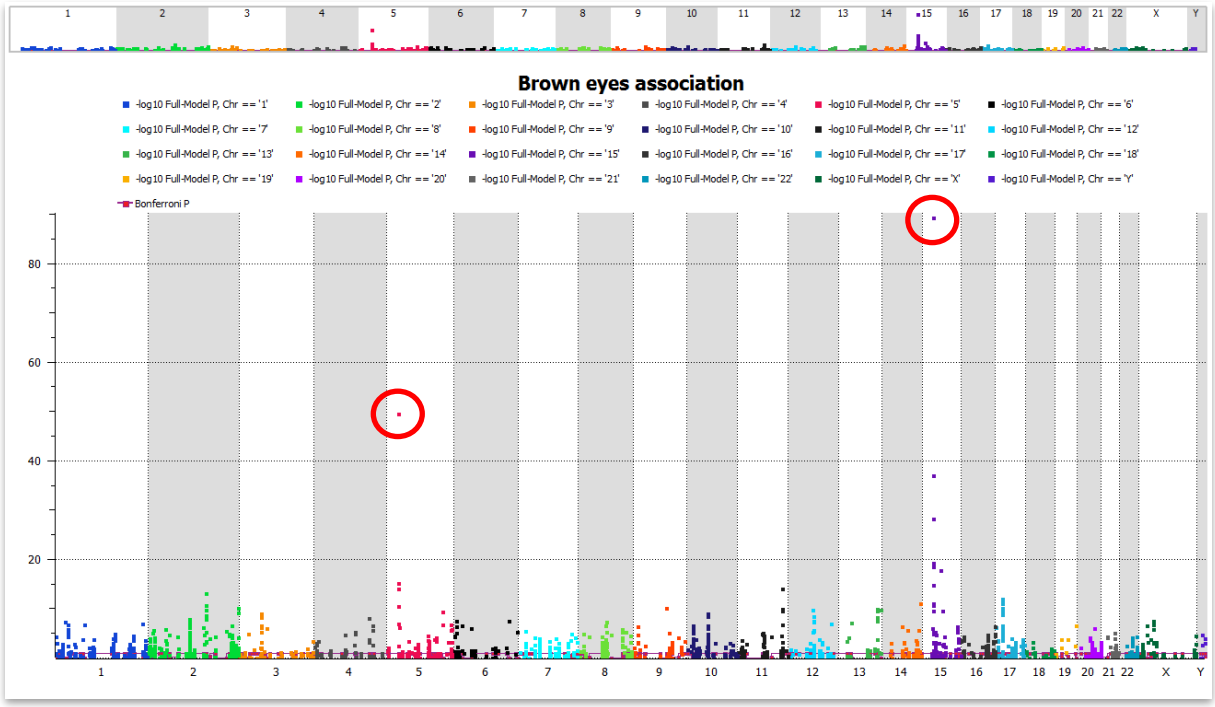


Figure 60. A Manhattan plot illustrating associations of genetic markers with brown eyes. Note the top associated SNPs: rs12913832 in HERC2 gene (chr. 15) and rs16891982 in SLC45A2 (chr. 5).

Table 45. A summary of genes significantly associated with eye colour, segregated into three groups according to their role in the pigmentation process. The genes are categorized according to their role in the pigmentation regulation. The relative protein role (or absence of such) in the pigmentation regulation has been inferred from the GeneCards web site: http://www.genecards.org/ and published references.

			Gene name	Role in pigmentation
Eye colour	Genes directly associated with pigmentation process		HERC2 SLC45A2 OCA2 SLC24A5 MYO5A SLC24A4	Major pigmentation predictors
			FLNB	A novel gene associated with skin pigmentation [385]
	'Novel' genes, previously not associated with pigmentation	Genes not associated directly with pigmentation	PCDH15	Expressed in the eye epithelium
			RPL6	Involved in retinal pigmentation and melanoma
			ARHGEF7	Possible involvement in melanoma
			POU2F3	Transcription factor in melanocytes and in keratinocyte differentiation [451]
			DDX3Y	Expressed in melanoma
		Novel genes previously not shown to be involved in the pigmentation process	NTRK2, STK33P1, EPN2, HDAC4, EIF2C3, PCNPP3, RPSAP52, CACNB4	No significant role in pigmentation found yet
			EDAR	Hair morphology [448, 449]

5.3.5.2. Hair colour

Human hair colour varies from the light blond and red, all the way through light brown to dark black shade. Hair pigmentation has been associated with at least 15 genes, with many genes playing a central role in this process, such as MC1R, ASIP, SLC45A2, OCA2, HERC2, SLC24A4, KITLG, TYR, TPCN2, TYRP1 and IRF4 [384, 389].

In the current study, various shades of hair pigmentation as well as hair morphology were tested for potential association with candidate genetic markers. Hair colour and hair morphology (very curly hair) association results are summarized in Table 46. Only the most significant results for the black hair are shown, as a very large number of significant markers (more than 250 SNPs in 78 genes) was identified.

After applying a 1.0E-5 p-value threshold, 48 SNPs in 19 genes were associated with different shades of hair colour (the black shade is represented by the top 30 markers) and 20 SNPs in 11 genes were found to be associated with curly hair (Table 46). Black hair association analysis demonstrated the highest number of significant SNPs (>250 markers), followed by brown (17 markers). Blond and red colours revealed a genetic association with five and two markers respectively. Notably, the p-value threshold applied in the hair colour association study was lower, compared to 1.0E-07 threshold, applied in the eye colour study, due to the higher number of significant markers associated with the latter phenotype.

Table 46. A summary of the hair colour and hair curliness association study.

Trait	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
Black hair	5:33951693	51.9114	rs16891982	SLC45A2	2.42E-67	4.96E-64
	15:28365618	63.5959	rs12913832	HERC2	8.38E-39	1.72E-35
	15:48426484	47.4004	rs1426654	SLC24A5	3.37E-31	6.91E-28
	17:19175317	65.7768	rs28591622	EPN2	6.88E-24	1.41E-20
	17:19204863	56.3708	rs4924980	EPN2	2.74E-22	5.62E-19
	11:120133494	98.3911	rs2715883	POU2F3	2.90E-22	5.95E-19
	17:19239432	42.9114	rs1043809	EPN2	6.72E-22	1.38E-18
	17:19172505	37.4594	rs28760541	EPN2	8.01E-21	1.64E-17
	13:102821327	44.81	rs4771420	FGF14	6.94E-20	1.42E-16
	14:101142890	43.2491	rs730570	C14orf70	1.35E-19	2.76E-16
	12:66115201	52.0996	rs12300373	near RPSAP52	2.09E-19	4.29E-16
	2:152829657	91.5424	rs6721518	CACNB4	2.20E-19	4.50E-16
	13:102816760	59.7048	rs2607642	FGF14	2.32E-19	4.75E-16
	11:120130512	46.8339	rs1941411	POU2F3	6.02E-19	1.23E-15
	17:19224397	62.5221	rs6587216	EPN2	7.22E-19	1.48E-15
	11:120118498	74.0258	rs2847502	POU2F3	2.73E-18	5.59E-15
	5:33958910	65.0609	rs1010872	SLC45A2	1.73E-17	3.54E-14
	13:102824147	43.845	rs9557826	FGF14	2.50E-17	5.13E-14
	17:19211073	40.4834	rs8072587	EPN2	3.71E-17	7.60E-14
	13:111827167	87.3155	rs9522149	ARHGEF7	1.14E-16	2.34E-13
	6:145055331	52.5387	rs4463276	UTRN	1.60E-16	3.29E-13
	11:120107411	46.9889	rs882856	POU2F3	2.02E-16	4.15E-13
	15:28268990	43.7915	rs749846	OCA2	2.06E-16	4.22E-13
	17:19247075	43.8229	rs4924987	B9D1	2.18E-16	4.46E-13
	5:33954511	80.2694	rs2287949	SLC45A2	2.19E-16	4.48E-13
	15:28516084	53.155	rs8039195	HERC2	2.53E-16	5.19E-13
	12:66168151	69.1679	rs7134682	-	2.92E-16	5.98E-13
	2:152829626	92.0793	rs16830527	CACNB4	3.67E-16	7.52E-13
	2:240282208	86.7675	rs12471054	HDAC4	4.41E-16	9.04E-13
	1:36367780	89.7804	rs595961	EIF2C1	5.49E-16	1.12E-12
Brown hair	5:33951693	51.9114	rs16891982	SLC45A2	1.30E-08	2.65E-05
	17:19204863	56.3708	rs4924980	EPN2	1.11E-07	2.27E-04
	17:19175317	65.7768	rs28591622	EPN2	1.80E-07	3.66E-04
	17:19239432	42.9114	rs1043809	EPN2	7.05E-07	1.44E-03
	17:19172505	37.4594	rs28760541	EPN2	1.16E-06	2.36E-03
	17:19224397	62.5221	rs6587216	EPN2	3.87E-06	7.89E-03
	15:48426484	47.4004	rs1426654	SLC24A5	9.08E-06	1.85E-02
	17:19211073	40.4834	rs8072587	EPN2	9.65E-06	1.97E-02
	15:28356859	69.3697	rs1129038	HERC2	1.75E-05	3.57E-02
	17:48168877	69.6587	rs11657072		1.78E-05	3.63E-02
	13:102821327	44.81	rs4771420	FGF14	2.07E-05	4.21E-02

Trait	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
Blond hair	15:28365618	63.5959	rs12913832	HERC2	2.62E-11	5.34E-08
	15:28356859	69.3697	rs1129038	HERC2	1.46E-06	2.98E-03
	15:28513364	50.3991	rs916977	HERC2	9.19E-06	1.87E-02
	14:92781001	44.607	rs4904868	near SLC24A4	1.29E-05	2.62E-02
	6:396321	41.821	rs12203592	IRF4	1.70E-05	3.47E-02
Red hair	8:4190796	64.2804	rs117368188	CSMD1	3.07E-06	6.25E-03
	5:79085726	95.0018	rs12657828	CMYA5	1.41E-05	2.88E-02
Very curly hair	8:116670666	58.3266	rs10505261	TRPS1	1.47E-23	3.00E-20
	8:116700847	40.9004	rs7842702	TRPS1	2.12E-21	4.33E-18
	2:232198327	74.1716	rs6719593	ARMC9	4.04E-21	8.23E-18
	7:41986689	82.1162	rs7789366	near GLI3	6.54E-21	1.33E-17
	8:72131359	68.3893	rs73684719	EYA1	2.80E-20	5.71E-17
	4:5638652	69.3266	rs4353849	EVC2	2.95E-20	6.01E-17
	8:116776330	42.4465	rs10505268	TRPS1	3.92E-20	7.98E-17
	7:107696289	64.7565	rs17154865	LAMB4	1.32E-19	2.69E-16
	2:232198375	73.4834	rs6747710	ARMC9	4.98E-19	1.01E-15
	20:45801340	61.5517	rs6090632	EYA2	1.13E-17	2.29E-14
	8:116673229	40.3579	rs10955754	TRPS1	1.24E-17	2.52E-14
	7:107693455	46.6052	rs11487091	LAMB4	2.16E-17	4.41E-14
	8:116823046	51.3579	rs6981915	near EIF3H	1.33E-16	2.71E-13
	7:107704688	66.2269	rs10257477	LAMB4	3.10E-16	6.33E-13
	4:5638799	24.6458	rs10001971	EVC2	1.44E-15	2.93E-12
	5:138439801	69.655	rs1368374	SIL1	3.86E-15	7.87E-12
	7:83033303	34.321	rs2722980	SEMA3E	8.90E-15	1.81E-11
	7:83030169	26.3506	rs2709927	SEMA3E	1.17E-14	2.39E-11
	7:83041344	60.8137	rs2709963	SEMA3E	1.30E-14	2.65E-11
	22:46764108	59.5018	rs73448947	CELSR1	4.79E-14	9.77E-11

Table 47 summarizes the hair colour and morphology association results segregated according to each gene's role in pigmentation regulation. The electronic version of this table includes web links to the relevant references.

Table 47. A summary of genes, associated with hair colour and curly hair according to their role in the pigmentation regulation. The genes are categorized according to their role in the pigmentation regulation. The relative protein role (or absence of such) in the pigmentation regulation has been inferred from the GeneCards web site: http://www.genecards.org/ and published references.

			Gene name	Role in pigmentation
Hair pigmentation	Genes directly associated with hair colour		SLC45A2, HERC2, SLC24A5, OCA2, SLC24A4, IRF4	Major genes effecting pigmentation pattern [390, 394, 452]
	'Novel' genes, previously not associated with pigmentation	Genes indirectly associated with hair pigmentation	FGF14 ARHGEF7 HDAC4 POU2F3	Expressed in developing retina Risk factor in melanoma Involved in the melanin synthesis Transcription factor in melanocytes
		Novel genes previously not shown to be involved in the pigmentation process	EPN2, C14orf70, RPSAP52, CACNB4, UTRN, B9D1, EIF2C1, CSMD1, CMYA5, GLI3, EYA1, EVC2, EIF3H, SIL1, SEMA3E, CELSR1	No significant role in pigmentation found yet
Very curly hair			TRPS1 ARMC9 GLI3 EYA2 LAMB4 TRPS1, EIF3H	Hair morphogenesis Melanoma biomarker Essential for hair follicle development Promotes hair cell fate Downregulated in melanoma Associated with Langer-Giedion syndrome - symptoms include sparse scalp hair, thin upper lip and rounded nose [453]

A number of significantly associated genes such as SLC45A2, HERC2, SLC24A5, OCA2, SLC24A4, IRF4 were previously shown as important pigmentation regulators, while three (3) SNPs (rs16891982, rs12913832 and rs1426654) were described as the main predictors of hair colour (see Figure 61), as discussed in Section 5.1.2 and reviewed elsewhere [14, 72].

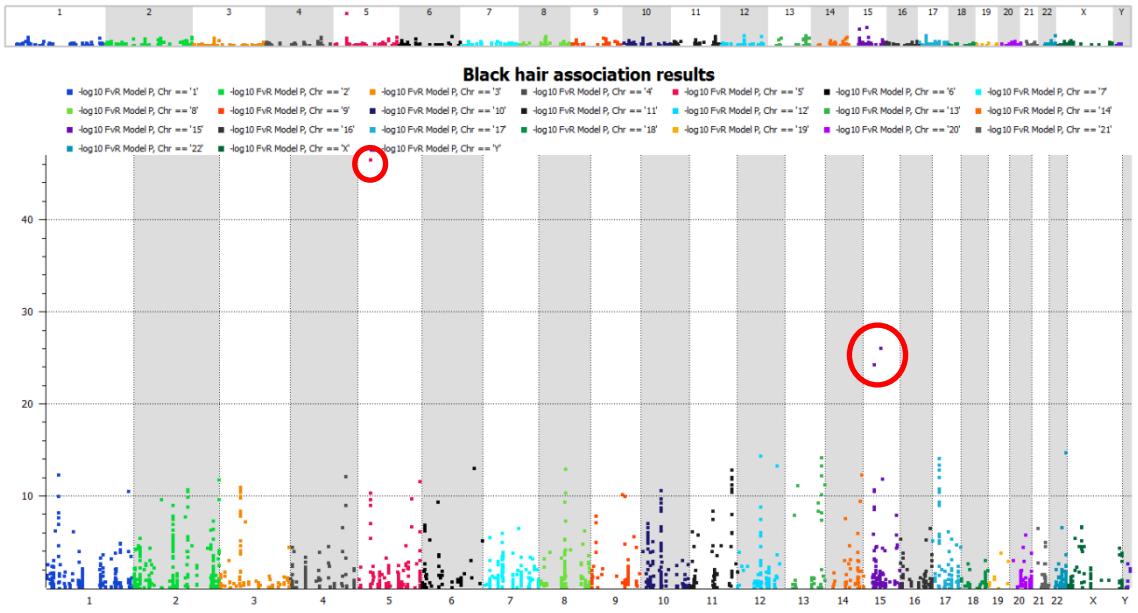


Figure 61. Manhattan plot illustrating associations of genetic markers with black hair. The most highly associated SNPs are highlighted in red (rs16891982 on chromosome 5) and purple (rs12913832 and rs1426654 on chromosome 15).

Several additional significant genes, such as FGF14, ARHGEF7, HDAC4 and POU2F3 were previously shown to be involved in various cellular processes, including hair morphogenesis, melanin synthesis and melanoma expression, albeit not associated directly with prediction of hair colour. Sixteen additional genes however, were not identified as pigmentation regulators both directly or indirectly and may represent novel candidates, affecting skin pigmentation.

Red hair did not produce any associations with markers in the MC1R genes, which is known as the major predictor of red hair [294, 454]. This was expected, as the number of available markers in this gene was greatly reduced due to manufacturer primer design failure (as discussed in Section 3.4.3). Nevertheless, red hair was associated with two markers in the CSMD1 and CMYA5 genes. Interestingly, the expression of CSMD1 is

localized at the top of sensory hair cells in the inner ear, although with no link to the body hair pigmentation [455]. The CMYA5 gene was hypothesized as a skeletal muscle regeneration regulator, with no known connection to pigmentation regulation as yet [323].

Markers in six genes, TRPS1, ARMC9, GLI3, EYA2, LAMB4 and EIF3H, showed significant association with very curly hair. All these genes are involved in either hair growth and morphology or are differentially expressed in melanoma. Notably, five of the six genes (not EIF3H), were significantly associated with African ancestry (as detailed in Section 5.3.4). This overlap is not surprising, given the hair morphology in the African population, although the association with curly hair is probably the primary one, providing indirect information on ancestry. Interestingly, the TRPS1 gene was previously associated with Hypertrichosis (also called Ambras syndrome). The major symptom of this disorder is an abnormal amount of hair growth over the body [450]. Additionally, the TRPS1 and another significantly associated gene EIF3H, were previously found to be involved in the Langer-Giedion syndrome. The symptoms of this disorder include sparse, slowly growing scalp hair and craniofacial dysmorphisms, such as bulbous-shaped nose, thin upper lip and large prominent ears [453]. The association of these six genes with very curly hair is novel and has not been described previously. Interestingly, the association of polymorphisms in the TRPS1 gene with nose morphology (specifically with nose width) was also demonstrated in the current study, as discussed in Section 5.3.6.

5.3.5.3. Skin colour

Normal human skin pigmentation ranges from very pale (fair) to very dark (black) shade, increasing in pigmentation towards the equator and providing better UV protection [72, 371]. With increased UV exposure, melanocytes produce more melanin, which is used by keratinocytes to protect DNA in the nuclei from UV damage. This complex process is regulated through signalling cascades of various factors and not fully understood yet.

This study analysed four shades of skin colour (fair, average, olive and dark) as well as freckling for potential association with genetic markers (see Table 48). The black skin genetic association analysis demonstrated the highest number of significantly associated markers with 367 markers when applying a 1.0E-7 threshold in approximately 100 genes and intergenic regions. The olive skin genetic association analysis revealed 33 significant markers in 13 genes and intergenic regions. The fair skin was associated with seven SNPs in six different genes, while the average shade did not produce any significant associations with genetic markers.

Table 48. A summary of the association study for skin colour and freckling.

Trait	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
Fair skin	3:58112440	56.6937	rs2362903	FLNB	9.44E-08	1.92E-04
	5:33951693	51.9114	rs16891982	SLC45A2	1.67E-07	3.40E-04
	6:396321	41.821	rs12203592	IRF4	3.08E-06	6.27E-03
	6:542416	84.7417	rs3799296	EXOC2	3.56E-06	7.26E-03
	15:28365618	63.5959	rs12913832	HERC2	5.42E-06	1.10E-02
	5:33969628	48.3911	rs35414	SLC45A2	1.25E-05	2.54E-02
	15:28344238	64.7122	rs7495174	OCA2	2.04E-05	4.15E-02
Olive skin	22:46835000	99.9133	rs4823810	CELSR1	2.59E-14	5.28E-11
	11:120170030	43.5	rs11217807	POU2F3	2.81E-14	5.72E-11
	13:34864240	64.5738	rs2065982	between RFC3 and NBEA	1.05E-13	2.14E-10
	11:120169962	41.0203	rs11217806	POU2F3	1.25E-13	2.56E-10
	11:120129533	55.4446	rs11217777	POU2F3	2.36E-12	4.82E-09
	12:66114478	77.5812	rs10784490	near RPSAP52	8.93E-12	1.82E-08
	12:112843363	49.4852	rs2301723	RPL6	1.63E-11	3.33E-08
	1:36099200	80.6328	rs6668101	PSMB2	2.73E-10	5.56E-07
	6:21911616	61.7712	rs7745461	-	7.38E-09	1.50E-05

	2:240282208	86.7675	rs12471054	HDAC4	8.76E-09	1.78E-05
	6:21911578	59.6015	rs6456456	-	1.27E-08	2.58E-05
	2:240295613	59.3506	rs2176046	HDAC4	1.72E-08	3.50E-05
	2:223089431	32.214	rs2289266	PAX3	2.98E-08	6.08E-05
	1:36437991	73.6494	rs12096413	EIF2C3	3.43E-08	7.00E-05
	2:109556418	48.8413	rs7598206	EDAR	5.43E-08	1.11E-04
	15:28344238	64.7122	rs7495174	OCA2	7.68E-08	1.56E-04
	2:109556365	49.2528	rs6542787	EDAR	9.21E-08	1.88E-04
	2:109616376	35.0627	rs17034770	near EDAR	1.11E-07	2.26E-04
	2:223086976	80.0074	rs6741337	PAX3	1.53E-07	3.12E-04
	2:109556761	63.1882	rs260710	EDAR	1.66E-07	3.38E-04
	11:120121323	45.8118	rs11217775	POU2F3	2.51E-07	5.12E-04
	2:16795905	67.9041	rs6724022	FAM49A	2.93E-07	5.97E-04
	2:109556667	66.5683	rs260709	EDAR	3.10E-07	6.32E-04
	5:174135737	46.5166	rs17063871	-	3.22E-07	6.55E-04
	2:109550092	54.8284	rs6761501	EDAR	3.29E-07	6.70E-04
	15:28268990	43.7915	rs749846	OCA2	3.62E-07	7.38E-04
	15:26905021	56.0904	rs17738087	GABRB3	4.04E-07	8.24E-04
	2:109557099	54.6218	rs260711	EDAR	4.26E-07	8.69E-04
	2:109599256	45.5572	rs260674	EDAR	5.31E-07	1.08E-03
	2:109586371	63.5959	rs11123719	EDAR	5.91E-07	1.20E-03
Dark skin	8:116700847	40.9004	rs7842702	TRPS1	7.15E-67	1.46E-63
	2:232198375	73.4834	rs6747710	ARMC9	5.44E-62	1.11E-58
	7:41986689	82.1162	rs7789366	near GLI3	6.67E-62	1.36E-58
	2:177992107	52.7546	rs6433650	near RPL29P8	5.24E-59	1.07E-55
	8:116776330	42.4465	rs10505268	TRPS1	1.75E-53	3.57E-50
	10:34755348	32.3266	rs1978806	PARD3	1.76E-53	3.58E-50
	8:116670666	58.3266	rs10505261	TRPS1	1.75E-52	3.57E-49
	7:83030169	26.3506	rs2709927	SEMA3E	5.67E-51	1.15E-47
	20:62164263	51.3413	rs6010957	PTK6	8.10E-50	1.65E-46
	6:137345768	34.5203	rs276488	IL20RA	1.06E-49	2.16E-46
	7:83033303	34.321	rs2722980	SEMA3E	8.28E-48	1.69E-44
	2:232198327	74.1716	rs6719593	ARMC9	8.09E-47	1.65E-43
	18:67672106	58.8616	rs56293475	RTTN	5.98E-46	1.22E-42
	7:107704688	66.2269	rs10257477	LAMB4	7.31E-46	1.49E-42
	7:107696289	64.7565	rs17154865	LAMB4	4.02E-45	8.19E-42
	8:116673229	40.3579	rs10955754	TRPS1	1.17E-43	2.39E-40
	8:72127562	34.5258	rs79867447	EYA1	7.12E-41	1.45E-37
	20:45801340	61.5517	rs6090632	EYA2	2.61E-40	5.32E-37
	8:72131359	68.3893	rs73684719	EYA1	6.05E-39	1.23E-35
	18:19765739	45.5203	rs16964572	GATA6	8.02E-38	1.63E-34
	7:83041344	60.8137	rs2709963	SEMA3E	8.83E-37	1.80E-33
	10:24195798	26.6513	rs73604433	KIAA1217	1.59E-36	3.23E-33
	4:5638652	69.3266	rs4353849	EVC2	3.96E-35	8.07E-32
	7:83041372	59.6753	rs2709964	SEMA3E	4.10E-35	8.36E-32
	7:107689844	57.6531	rs10260756	LAMB4	1.49E-34	3.03E-31

	5:78190471	60.5009	rs16876062	ARSB	1.34E-33	2.73E-30
	20:62164242	51.0535	rs6011878	PTK6	5.17E-33	1.05E-29
	12:56603834	42.6176	rs773658	RNF41	2.41E-32	4.92E-29
	7:107693455	46.6052	rs11487091	LAMB4	2.00E-31	4.07E-28
	17:72434959	48.6771	rs7222873	GPRC5C	3.16E-31	6.45E-28
Freckling	6:396321	41.821	rs12203592	IRF4	2.40E-10	4.89E-07

Fair skin genetic association with the HERC2, OCA2, SLC45A2, IRF4 and FLNB genes confirmed previously published results, as discussed in Section 5.1.2. Although the EXOC2 gene was known previously to be associated with hair colour, its association with skin pigmentation was novel in this current study [386]. The absence of significant markers associated with the average colour may be a result of an "ambiguous" definition of this category, as discussed in Section 5.3.1.2.

The olive and black skin shades produced numerous associations with eight (8) and eighteen (18) genes and intergenic regions respectively. While the TRPS1 and OCA2 genes are known to have a major impact on pigmentation and specifically melanin synthesis, both were not associated specifically with dark or olive skin. According to the literature, the main predictor of dark skin is rs6119471 in the agouti signalling protein (ASIP) gene [10, 456]. However, this SNP was not included in the current study due to technical difficulties with primer design. The most significantly associated genes with olive skin were the CELSR1 and POU2F3, while the TRPS1 and ARMC9 demonstrated association with dark skin. Both the TRPS1 and POU2F3 genes were previously associated with pigmentation, however both the CELSR1, ARMC9 and other significantly associated genes were not documented to be linked with pigmentation genetics. In the contrary, these two genes were found associated with craniofacial embryogenetics, including several associations with craniofacial measurements in this project (Section 5.3.6). On the one hand, these results may reveal a new functional role for these genes and their encoded proteins. On the other hand, this may be a somehow spurious result, because of the small sample size used for this specific trait analysis (olive and dark skin) and therefore, must be checked for replication in the future study.

In summary, eleven (11) genes associated with skin colour in this study were not previously described as pigmentation predictors, although they were linked to melanogenesis regulation in several studies (summarized in Table 49). Interestingly, one of these genes, LAMB4, was found to be downregulated in skin exposed to UV radiation and together with several other genes was included in an assay used for the detection of UV exposure [457].

SNP rs12203592, located in the enhancer region of the IRF4 gene, was found associated with freckling and fair skin. Notably, the association of rs12203592 in the IRF4 gene with freckling has replicated previously reported evidence of this gene's involvement in melanin synthesis [384, 458]. A recent study has found an association of this marker with high sensitivity to UV exposure, blue eyes and brown hair. It was demonstrated to regulate transcription of the TYR gene through interaction with TFAP2A and MITF transcriptional factors [459].

The summary of genes association with the three skin shades and freckling is summarised in Table 49. The electronic version of this table includes web links to the relevant references.

Table 49. A summary of genes, significantly associated with skin colour and freckling. The genes are categorized according to their role in the pigmentation regulation. The relative protein role (or absence of such) in the pigmentation regulation has been inferred from the GeneCards web site: http://www.genecards.org/ and published references.

			Gene name	Role in skin pigmentation
Skin pigmentation	Genes directly associated with skin colour		SLC45A2, HERC2, OCA2, TRPS1, IRF4	Major genes effecting pigmentation pattern [390, 394, 452]
			FLNB	A gene recently associated with skin pigmentation [385]
	'Novel' genes, previously not associated with pigmentation	Genes indirectly associated with pigmentation	POU2F3	Transcription factor in melanocytes and keranocytes
			RPL6	Differentially expressed in melanoma
			HDAC4	Expressed in melanocytes
			PAX3	Involved in melanogenesis
			ARMC9	Up-regulation in melanoma
			GLI3	Required for retinal pigment development
			PARD3	Increased expression in melanoma
			SEMA3E	Potentially involved in retinal pigment epithelial cell activity
			PTK6	Potentially involved in melanoma
			IL20RA	Deregulated in melanoma
			LAMB4	Differentially regulated in exposure to UV radiation
		Novel genes previously not shown to be involved in the pigmentation process	CELSR1, RFC3, PCNPP3, PSMB2, EIF2C3, EDAR, FAM49A, GABRB3, RPL29P8, RTTN, EYA1, EYA2, GATA6, KIAA1217, EVC2, ARSB, RNF41, GPRC5C, NBEA, RPSAP52	No significant role in skin pigmentation identified yet

The association analysis of eye, skin and hair colour produced a large number of strongly associated markers. The most significantly associated SNPs were in full concordance with previously published studies, confirming statistical methods used in this project and providing a solid basis for subsequent craniofacial traits analysis. In

addition, a significant number of novel genes either not associated with pigmentation or not associated with a specific trait previously, were identified. This information will extend the current knowledge of the pigmentation genetics and assist in developing more accurate forensic assays for eye, skin and hair colour prediction.

5.3.6. Craniofacial traits association study

This section summarises the results of the craniofacial traits genetic association study. Over 100 craniofacial traits were analysed for potential association with genetic markers. The association analysis was performed using a stringent linear regression model, incorporating a correction factor for population stratification and covariates such as sex and BMI. The population stratification issue was discussed in Sections 5.2.5 and 5.3.2. The use of covariates in the statistical analysis aimed to reduce the risk of introducing confounding effects, which can result in false positive associations. While sexual dimorphism in the craniofacial morphology is a well-known aspect [460-462], BMI may also affect craniofacial traits, since the soft facial tissue may change significantly with weight gain or loss. Despite that, this potential confounding factor has been disregarded in association studies of normal craniofacial morphology performed to date. Age was not considered a significant covariate, given that almost all the volunteers participated in this study were young students whose average age was approximately 27 years old.

The results of the association analyses of the craniofacial traits are summarized in Tables 50-53. In general, thirteen linear distances, four angular distances and one ratio revealed significant associations with 58 genetic markers in 33 different genes and intergenic regions (Tables 50 and 51). The nasal area measurements, including either “n”, “prn”, “sn” or “al” landmarks, produced the majority of the total number of significant associations. The superior reproducibility of the nasal area landmarks compared to other landmarks (as discussed in Section 4.4) is likely a factor in these findings. Additional measurements such as maximum facial width, upper lip width and ear height also produced significant associations with various genetic markers.

An association analysis of a non-anthropometrical phenotypic trait - single vs double eye lid, demonstrated significant association with nine markers in six genes and intergenic regions (Table 52).

The most significant associations with genetic markers (based on p-values) were produced by the principal components (Table 53). This result is not surprising, as the principal components represented vectors of different craniofacial measurements and as a result were expected to produce more significant associations with SNPs (both p-value and number–wise). The 3D craniofacial measurements association results (Section 5.3.6.1) are followed by the principal components association results and discussion (Section 5.3.6.2). Overall, the statistical analysis revealed 161 unique markers in 63 genes and intergenic regions associated with either craniofacial measurements or principal components.

Table 50. A summary of the association analyses of the linear craniofacial distances.

Trait	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
Eu-Eu	1:103420759	47.8524	rs4908280	COL11A1	4.60E-07	9.26E-04
	1:103483514	41.4649	rs945748	COL11A1	8.13E-07	1.64E-03
	1:103444679	61.4668	rs11164649	COL11A1	5.08E-06	1.02E-02
	2:121708589	76.0554	rs60613333	GLI2	1.64E-05	3.31E-02
Cephalic index	2:121708589	76.0554	rs60613333	GLI2	1.16E-07	2.35E-04
	2:121708596	77.393	rs60640022	GLI2	1.57E-07	3.17E-04
	2:121708140	47.8524	rs11899735	GLI2	1.88E-06	3.80E-03
	1:36099200	80.6328	rs6668101	PSMB2	1.27E-05	2.56E-02
	4:5638652	69.3266	rs4353849	EVC2	1.67E-05	3.38E-02
	6:44405327	37.19	rs559035	CDC5L	2.09E-05	4.21E-02
	1:103444679	61.4668	rs11164649	COL11A1	2.24E-05	4.51E-02
zy-zy	16:71186860	133.7119	rs3114614	HYDIN	6.76E-06	1.36E-02
al-al	1:156098620	66.3467	rs528636	LMNA	3.31E-08	6.66E-05
	1:156106185	80.0779	rs505058	LMNA	1.81E-07	3.65E-04
	14:23598760	44.9446	rs7151708	SLC7A8	1.32E-06	2.65E-03
	1:156100739	63.1081	rs61726477	LMNA	5.70E-06	1.15E-02
	10:34755348	32.3266	rs1978806	PARD3	1.49E-05	3.00E-02
	2:232198375	73.4834	rs6747710	ARMC9	1.78E-05	3.59E-02
	8:116776330	42.4465	rs10505268	TRPS1	2.04E-05	4.12E-02
g-gn	6:137345768	34.5203	rs276488	IL20RA	4.13E-06	8.30E-03
	7:19129767	51.8376	rs11981706	TWIST1	2.54E-05	5.10E-02
	7:19104204	102.6494	rs2717340	TWIST1	6.25E-05	1.26E-01
n-sto	1:36367780	89.7804	rs595961	EIF2C1	7.90E-07	1.59E-03
	15:26905021	56.0904	rs17738087	GABRB3	1.02E-06	2.04E-03

n-prn	14:101142890	43.2491	rs730570	C14orf70	3.84E-06	7.71E-03
	22:35466515	56.9472	rs362038	ISX	9.36E-05	1.88E-01
n-sn	15:26905021	56.0904	rs17738087	GABRB3	2.55E-08	5.12E-05
	1:79225423	84.8023	rs1937025	-	7.73E-06	1.55E-02
	1:36367780	89.7804	rs595961	EIF2C1	1.11E-05	2.23E-02
	6:145055331	52.5387	rs4463276	UTRN	1.62E-05	3.26E-02
	22:35470453	138.2841	rs1883384	ISX	1.74E-05	3.50E-02
sl-gn	7:55131686	38.6402	rs17335912	EGFR	9.20E-06	1.85E-02
	5:112572574	58.1162	rs2416309	MCC	1.94E-05	3.90E-02
prn-go	1:197109042	39.2657	rs1332663	ASPM	9.50E-08	1.91E-04
	1:197070815	45.7786	rs1412640	ASPM	2.22E-07	4.46E-04
	1:197059892	24.9428	rs10754214	ASPM	5.80E-07	1.17E-03
	1:197086669	48.786	rs10754215	ASPM	7.67E-07	1.54E-03
g-pg	6:137345768	34.5203	rs276488	IL20RA	3.24E-05	6.53E-02
sa-sba	2:152814028	45.7177	rs16830498	CACNB4	1.89E-05	3.80E-02
	20:62164263	51.3413	rs6010957	PTK6	4.72E-05	9.49E-02
ls-sto	2:109579738	62.7011	rs260690	EDAR	8.41E-08	1.69E-04
	2:109562495	67.6753	rs260714	EDAR	5.43E-07	1.09E-03
	13:111827167	87.3155	rs9522149	ARHGEF7	1.31E-06	2.64E-03
	2:109567925	64.8321	rs260700	EDAR	1.39E-06	2.79E-03
	12:66168151	69.1679	rs7134682		1.74E-06	3.51E-03
	1:36367780	89.7804	rs595961	EIF2C1	1.90E-06	3.82E-03
	13:102816760	59.7048	rs2607642	FGF14	1.91E-06	3.84E-03

Table 51. A summary of the association analyses of the angular craniofacial distances and ratios between the linear distances.

	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
nasal tip protrusion/nose height index sn-prn)x100/(n-sn	17:19204863	56.3708	rs4924980	EPN2	8.25E-07	1.66E-03
	6:145055331	52.5387	rs4463276	UTRN	1.51E-06	3.04E-03
	17:19175317	65.7768	rs28591622	EPN2	2.11E-06	4.24E-03
	17:19239432	42.9114	rs1043809	EPN2	3.50E-06	7.03E-03
	17:69512099	65.2251	rs10512572	-	5.88E-06	1.18E-02
	17:19211073	40.4834	rs8072587	EPN2	1.35E-05	2.72E-02
	9:12704725	54.3616	rs2733832	TYRP1	1.36E-05	2.73E-02
	17:19224397	62.5221	rs6587216	EPN2	1.62E-05	3.25E-02
	17:19172505	37.4594	rs28760541	EPN2	2.33E-05	4.69E-02
nasal tip angle (n-prn-sn)	7:107696289	64.7565	rs17154865	LAMB4	7.98E-06	1.60E-02
	16:86386367	37.7712	rs11642858	-	1.70E-05	3.42E-02
nasal vertical prominence angle (tr-prn-gn)	15:26905021	56.0904	rs17738087	GABRB3	2.51E-06	5.04E-03
	2:109579738	62.7011	rs260690	EDAR	5.98E-06	1.20E-02
	2:109567925	64.8321	rs260700	EDAR	6.98E-06	1.40E-02
	2:109556761	63.1882	rs260710	EDAR	1.35E-05	2.71E-02
	1:234330571	58.7768	rs7531988	SLC35F3	1.66E-05	3.33E-02
	8:72264982	3.369	rs6988283	EYA1	1.79E-05	3.59E-02
	11:120126277	40.2825	rs2715881	POU2F3	2.34E-05	4.70E-02
nasolabial angle (prn-sn-ls)	4:146418167	34.762	rs17020235	SMAD1	1.91E-06	3.85E-03
nasofrontal angle (g-n-prn)	12:83421757	56.9631	rs10862564	TMTC2	3.73E-06	7.49E-03
	12:83421860	54.1384	rs10506886	TMTC2	3.85E-06	7.74E-03

Table 52. An association analysis of the eye lid (single/double).

Trait	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
eye lid	16:71186834	135.1328	rs3094869	HYDIN	1.15E-15	2.35E-12
	22:46835000	99.9133	rs4823810	CELSR1	1.55E-09	3.17E-06
	16:71186740	134.2694	rs3094868	HYDIN	2.04E-08	4.16E-05
	2:224794572	53.0295	rs4674831	WDFY1	3.09E-07	6.31E-04
	11:120170030	43.5	rs11217807	POU2F3	4.92E-07	1.01E-03
	16:71186860	133.7119	rs3114614	HYDIN	1.40E-06	2.86E-03
	20:55762947	60.5941	rs6123674	BMP7	1.42E-06	2.90E-03
	2:177418275	49.8133	rs725395	near MTX2	7.58E-06	1.55E-02
	11:120169962	41.0203	rs11217806	POU2F3	7.93E-06	1.62E-02

The analysis of direct cranial measurements revealed a significant association between genetic markers and both the Cephalic index (CI) and the maximum cranial width (eu-eu), which is used for the calculation of CI. There was also an overlap between the markers and genes in both traits. No significant associations of the maximum cranial length (g-op), also used for calculation of the CI, were found. This outcome may be a result of the insufficient representation of the candidate markers influencing this and possibly other traits.

The annotations of gene function in this section are based on the RefSeq database (http://www.ncbi.nlm.nih.gov/refseq/), GeneCards web portal (http://www.genecards.org) and AmiGO database (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi), while the information on the related disorders was sourced from the MalaCard database (http://www.malacards.org/) and OMIM database (http://omim.org/). SNPs annotation was performed using a Regulome database (http://www.regulomedb.org/) and Polyphen database (http://genetics.bwh.harvard.edu/pph2/) search, as detailed in the Material and Methods section.

5.3.6.1. Direct craniofacial measurements association results

The most significant gene found to be associated with the eu-eu distance as well as with the CI (although less significant), was collagen (COL11A1). Based on the RefSeq database information, this gene encodes one of the two alpha chains of type XI fibrillar collagen and is known to have multiple transcripts, as a result of an alternative splicing. The secreted protein is hypothesised to play an important role in fibrillogenesis by controlling lateral growth of collagen II fibrils. Interestingly, this gene was found mutated in Stickler Syndrome (OMIM: 604841) and Marshall Syndrome (OMIM: 154780) [99]. These two inherited disorders have very similar phenotypes and are characterized by distinctive facial appearance, such as flat midface, very small jaw, cleft lip/palate, large eyes, short upturned nose, eye abnormalities, round face and short stature. However, the facial features of Stickler syndrome are less severe and include a flat face with depressed nasal bridge and cheekbones, caused by underdeveloped bones in the middle of the face. Notably, another member of the collagen family, COL17A1 was recently associated with the distance between the eyeballs and the nasion [77]. This finding can be considered as a confirmation for this study result.

Another gene, GLI2, which was found in association with both eu-eu distance and CI, encodes a transcriptional factor that binds to DNA via zinc finger motifs. This gene is an activator of the Sonic Hedgehog (SHH) signalling pathway, which plays a central role in embryogenesis. The encoded protein is associated with cleft lip and palate as well as with other genetic disorders, such as Greig cephalopolysyndactyly syndrome, Pallister-Hall syndrome and preaxial and postaxial polydactyly.

Three additional genes that were found strongly associated with CI were PSMB2, EVC2 and CDC5L. The PSMB2 gene encodes a proteasome beta type subunit with a trypsin-like activity, although without any known involvement in embryonic craniofacial development. The EVC2 gene encodes a positive regulator of the hedgehog signalling pathway and plays a critical role in bone formation and skeletal development. The EVC2 gene was named after the Ellis Van Creveld Syndrome 2, which is associated with this gene [463]. This disorder is characterised by various skeletal abnormalities and especially by a very short stature. The ENTREZ database suggests that the cell division cycle 5-like (CDC5L) is a DNA-binding protein involved in the cell cycle control and forms an essential component of a non-snRNA spliceosome. The UniProtKB/Swiss-Prot web portal proposes that it may act as a transcription activator required for activating pre-mRNA splicing. However, GeneCards and MalaCards web site searches,

did not find any association of the PSMB2 and CDC5L genes with either craniofacial development or related inherited disorders.

Functional prediction of COL11A1 and GLI2 genes using the GeneMania web site, revealed a strong interaction between these factors (as well as with the COL17A1 gene, discussed above) through a network of genetic and physical interactions, as shown in Figure 62. A search with all five genes that were associated with the Cephalic index (including COL11A1 and GLI2 genes) also found a complex network of interactions between all these factors (Figure 63).

The results of this study provide important piece of evidence explaining, although partly, the complexity of processes affecting craniofacial embryonic development and expand the current knowledge in this field.

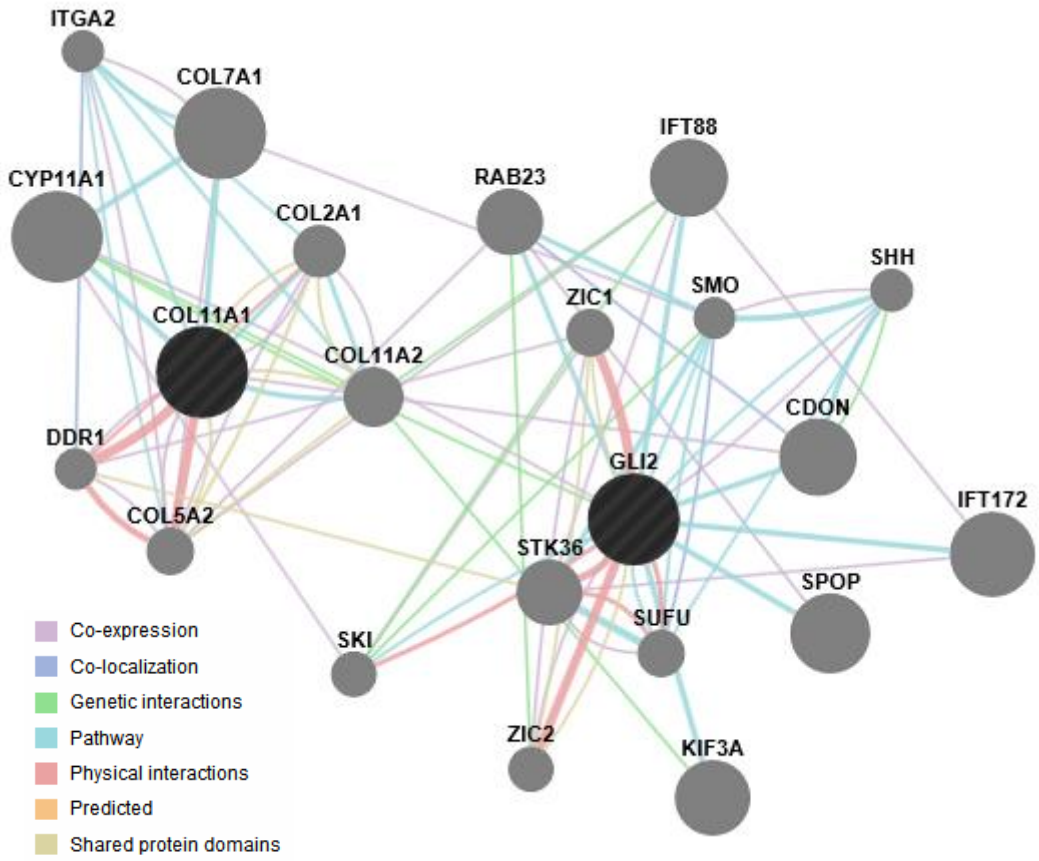


Figure 62. COL11A1 and GLI2 protein interactions based on the GeneMania database output (http://www.genemania.org/).

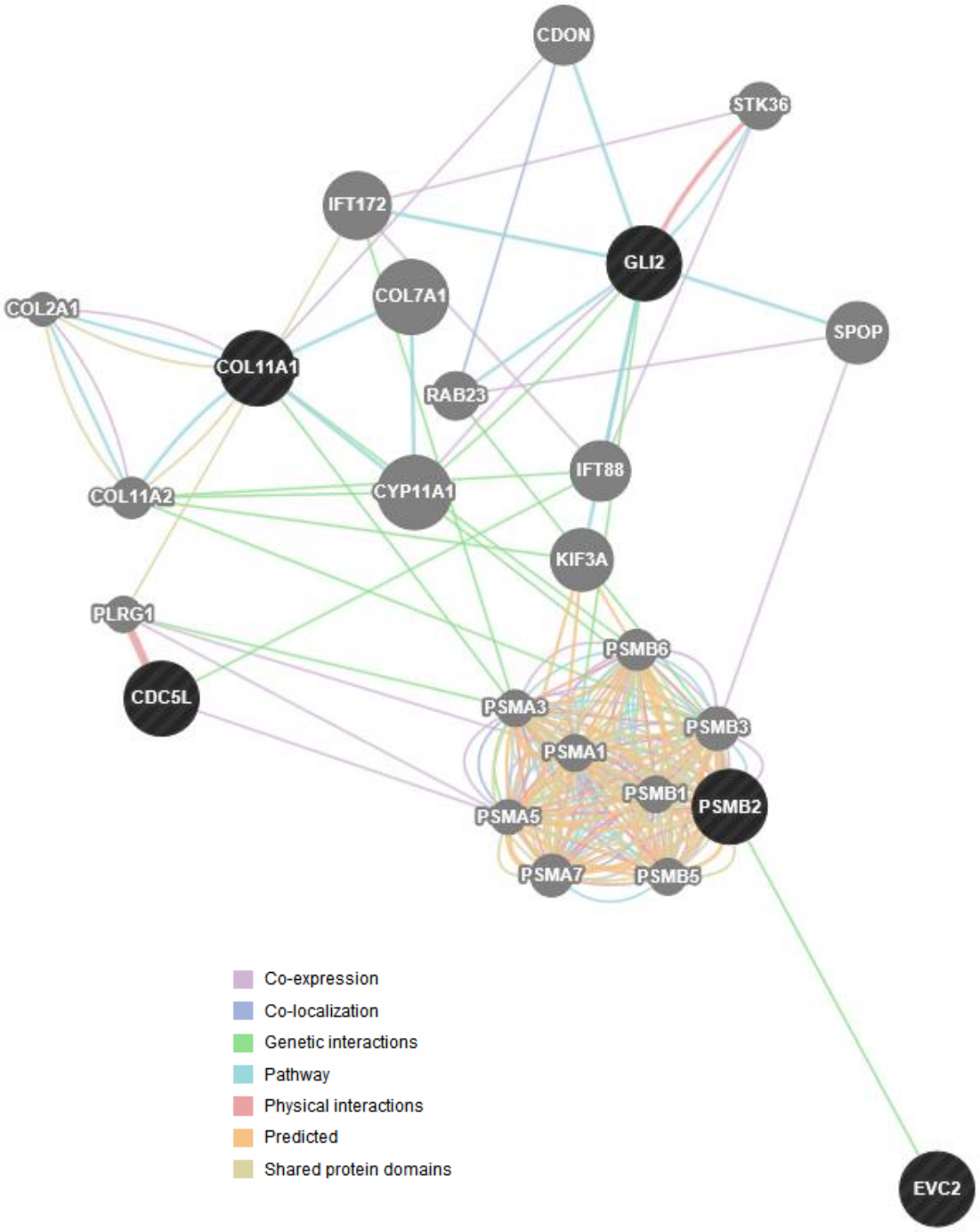


Figure 63. Interactions between factors found in significant association with the Cephalic index, based on the GeneMania web site output (http://www.genemania.org/).

Bizygomatic distance (maximum facial breadth) association results

Maximum facial breadth distance (zy-zy) produced a significant association with rs3114614, which is located in the intron of the Hydrocephalus inducing gene (HYDIN). This gene encodes a protein that is hypothesized to be involved in sperm cell cilia motility. However, it may have additional functions, given that mutations in HYDIN are associated with Primary Ciliary Dyskinesia 5 and Hydrocephalus (OMIM: 236600) disorders in both human and mouse models. The Hydrocephalus condition is characterized by a very large head size due to excessive accumulation of fluid in the brain, although the pathophysiology and biochemistry of this process is not clear.

While HYDIN has not previously been linked to normal craniofacial morphology, the bizygomatic distance was recently associated with SNP rs987525 near the CCDC26 gene, although with a relatively weak p-value of 0.017 [160]. Although not contradictory to the published results (as the CCDC26 gene was not covered in the current research due to manufacturer primer design failure), this study demonstrated a stronger evidence of association, based on the p-value of 6.76E-06.

Nasal area association results

Five linear morphological distances, four angular distances and one ratio, all from the nasal area demonstrated significant associations with twenty genes, including LMNA, EIF2C1, GABRB3, ASPM, PARD3, EPN2 and TRPS1. The following section provides a brief annotation of these genes function and description of their potential involvement in craniofacial disorders.

Lamin A (LMNA) together with other Lamin proteins, is a component of a fibrous layer on the nucleoplasmic side of the inner nuclear membrane, which provides a framework for the nuclear envelope and also interacts with chromatin. LMNA encoded protein acts to disrupt mitosis and induces DNA damage in vascular smooth muscle cells, leading to mitotic failure, genomic instability, and premature senescence of the cell. This gene has been found mutated in Mandibuloacral Dysplasia which is characterized by various skeletal and craniofacial abnormalities, including delayed closure of the cranial sutures and undersized jaw [464].

Argonaute RISC Catalytic Component (EIF2C1) is a member of a gene family, which plays a role in RNA interference. It is required specifically for the RNA-mediated gene

silencing and repression of translation by microRNAs. To date, this gene has not been linked to normal craniofacial development or to craniofacial malformations as to date.

The GABA-Alpha Receptor Beta-2 Subunit (GABRB3) gene encodes a subunit of a chloride channel that serves as the receptor for gamma-aminobutyric acid, a major inhibitory neurotransmitter of the nervous system. This gene was associated with the pathogenesis of several disorders including nonsyndromic orofacial clefts and Prader-Willi syndrome (OMIM:176270), which is characterised by specific facial appearance such as elongated face, thin upper lip and a prominent nose.

Abnormal Spindle-Like Microcephaly-Associated Protein (ASPM) is the human ortholog of the Drosophila melanogaster 'abnormal spindle' gene (asp), which is essential for normal mitotic spindle function in embryonic neuroblasts. Mutations in this gene are associated with microcephaly primary type 5.

Par-3 Family Cell Polarity Regulator protein (PARD3) is known to have multiple alternatively spliced transcript variants and was described to affect asymmetrical cell division and direct polarized cell growth. This gene may be involved in signalling pathways, directing cells through proliferation and migration in the developing face, which regulate facial symmetry.

Alternate splicing of the Epsin 2 (EPN2) gene results in multiple transcripts, encoding different isoforms, which play a role in the formation of clathrin-coated invaginations and endocytosis. A duplication/deletion of the 17p11.2 chromosomal region (including EPN2 gene) is associated with Smith–Magenis syndrome [465]. Interestingly, this syndrome is characterized by specific craniofacial features, including cleft lip/palate, microcephaly, triangular face and a broad nasal bridge.

Trichorhinophalangeal Syndrome I (TRPS1) encodes a transcription factor that binds specifically to GATA sequences and represses expression of GATA-regulated genes at selected stages of vertebrate development. Known to regulate chondrocyte proliferation and differentiation. Mutations in this gene were previously associated with Ambras syndrome (discussed in Sections 5.3.4 and 5.3.5.2) and with Langer-Giedion syndrome (OMIM: 190350). The symptoms of the latter syndrome include sparse, slowly growing scalp hair and craniofacial dysmorphisms, such as rounded nose, thin upper lip and large prominent ears [453]

A functional analysis of the 20 significant genes associated with nasal area measurements revealed that 19 of these factors (all but ASPM) interconnect with each other, forming a complex cluster of interactions on various molecular levels (Figure 64).

The Regulome annotation of the SNPs significantly associated with the linear and angular measurements, revealed that 36 of the 61 markers may be involved in various types of regulatory activity including affecting promoter regions and transcriptional binding sites.

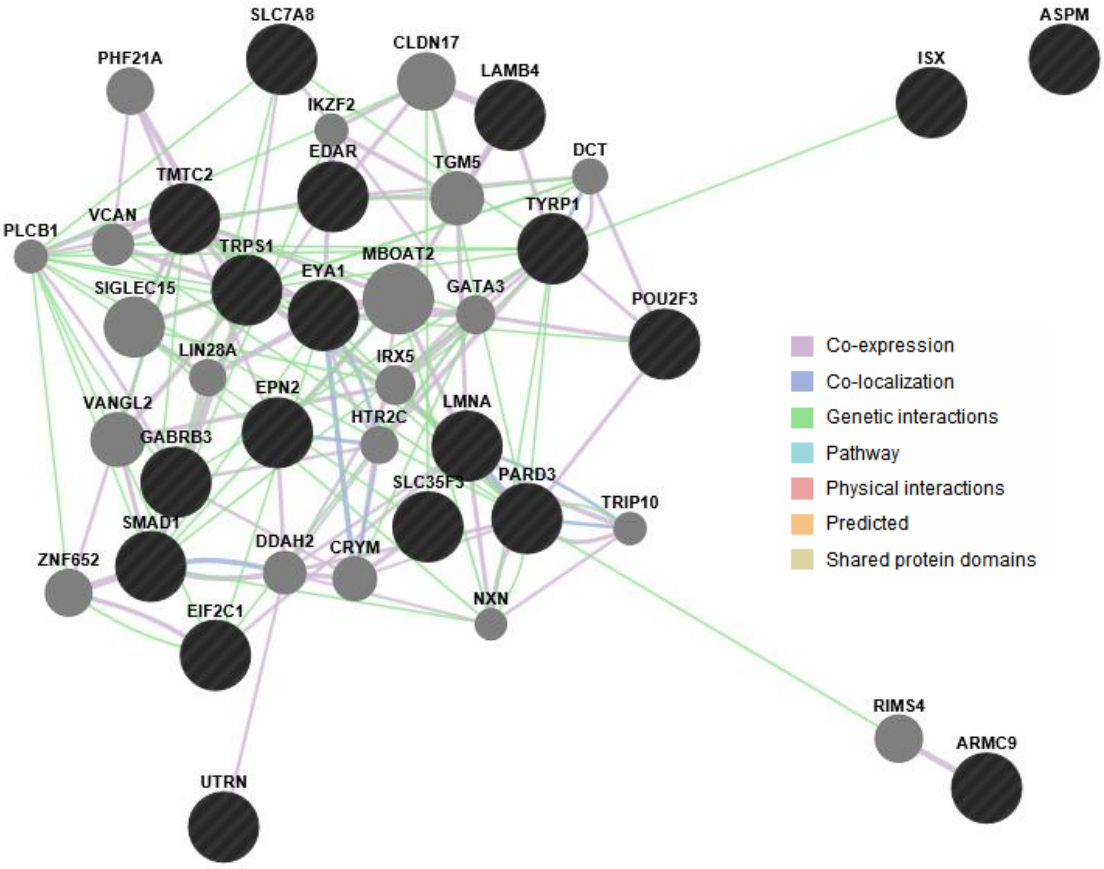


Figure 64. A network chart of genes, found in association with the nasal area measurements. Based on the GeneMania web site search (http://www.genemania.org/).

Eyelid genetic association results

Eyelid is a non-anthropometric, although a highly visible facial trait. The formation of eyelids in foetus starts at approximately 8 weeks, while the folds of surface ectoderm overgrow the eyes to form the eyelids, which remain closed until the seventh month of development. The eyelid can be largely segregated into double or single (according to presence or absence of a skin fold), while the former is more dominant in the East Asian population and the latter in the rest. The association analysis of this trait revealed a strong association with three SNPs located in the HYDIN gene. This gene was also associated with the bizygomatic distance in the current study. It might be hypothesized that the same gene is responsible for regulating cell migration and differentiation in different parts of the developing head, perhaps even in the forehead area specifically. Considering the high p-value of this association (1.15E-15), this marker might be a good predictor of this trait and provide indirect information on the East Asian ancestry.

Several other genes were also associated with the eyelid, such as CELSR1, POU2F3, WDFY1 and BMP7. The following section provides a brief overview of these genes and encoded protein function as well as their potential involvement in the craniofacial embryogenetics.

CELSR1 gene represents a part of the cadherin superfamily, involved in cell adhesion and receptor-ligand interactions. In general, cadherins participate in signal transduction in the WNT signalling pathway, which in central to embryonic development. This specific protein is a developmentally regulated neural-specific factor, which plays an unspecified role in early embryogenesis. AmiGO ontology database search found however, that this gene may play a role in the establishment of planar cell polarity. Specifically, it may be one of the several factors that coordinate organization of groups of cells in the plane of an epithelium, such that they all orient in the similar direction. Similarly to the PARD3 gene, discussed above, it may also play a role in regulating facial symmetry. Interestingly, the CELSR1 deficiency in mice results in various craniofacial defects, such as craniorachischisis (neural tube defects) and failure of the eyelid closure [466, 467]. Notably, CELSR1 was also found in significant association with the Asian ancestry in this project (Table 41) and can be used as an AIM .

POU2F3 gene encodes a transcriptional factor, which regulates cell type-specific differentiation pathways. The encoded protein is primarily expressed in the epidermis and plays a critical role in keratinocyte proliferation and differentiation, while also being a candidate tumour suppressor protein. In the current study, this gene was also

associated with European and Asian ancestry, brown eyes, black hair and olive skin as well as with nasal vertical prominence angle. The association of POU2F3 with the eyelid or any other craniofacial feature has not been reported before.

WD Repeat and FYVE Domain Containing 1 (WDFY1) encoded protein mediates the recruitment of proteins involved in membrane trafficking and cell signalling with localization in endosomes. To date, it has not been described as a craniofacial development factor.

Conversely, bone morphogenetic protein 7 (BMP7) encodes a growth factor. It plays a role in early development, specifically in bone inductive activity through calcium regulation and bone homeostasis. Deletion of this factor in mice leads to severe abnormalities of the orofacial complex [204].

Another significantly associated SNP rs725395 is an intergenic marker, which is located between MTX2 gene and MIR1246 gene. While both genes have not been associated with craniofacial morphology regulation, this marker is located in the transcriptional factor binding site (based on the Regulome search). Hypothetically, it may be a part of an enhancer element, which can be located far away from the regulated gene, while fine-tuning the craniofacial morphology variation indirectly, as has been proposed previously [213].

5.3.6.2. Principal components association results

Twenty principal components, representing all craniofacial measurements collected in this study (n=92), were generated using SVS software and analysed for potential association with candidate genetic markers. Of the twenty principal components that were generated, fourteen PCs were associated with 130 significantly associated markers in 57 different genes, with approximately half of genes common to 3D measurements and principal components (Table 53).

It should be noted however, that eigenvectors (hence principal components) analysed in this study, are not identical (although similar) to eigenvectors shown in Table 53 and Figure 54 as they were generated using all craniofacial measurements (and not only linear and angular distances). In addition, due to technical reasons, they were generated

using the SVS software, rather than PASW Statistics (as in Section 4.6), which theoretically may use slightly different algorithms for principal component analysis.

Table 53. An association analyses of principal components.

Trait	Marker	Average Depth	rsID	Gene Name	FvR Model P-Value	Bonferroni P-Value
PC1	16:71186860	133.7119	rs3114614	HYDIN	1.75E-15	3.54E-12
	16:71186834	135.1328	rs3094869	HYDIN	6.18E-12	1.25E-08
	16:71186888	112.1089	rs3094870	HYDIN	1.82E-10	3.68E-07
	16:71186740	134.2694	rs3094868	HYDIN	5.03E-09	1.02E-05
PC2	16:71186860	133.7119	rs3114614	HYDIN	3.02E-16	6.09E-13
	16:71186834	135.1328	rs3094869	HYDIN	9.45E-16	1.91E-12
	16:71186740	134.2694	rs3094868	HYDIN	3.46E-13	7.00E-10
	16:71186888	112.1089	rs3094870	HYDIN	2.67E-12	5.39E-09
	13:111827167	87.3155	rs9522149	ARHGEF7	4.77E-08	9.64E-05
	3:58006600	58.8173	rs839241	FLNB	6.89E-08	1.39E-04
	10:55949899	34.7048	rs16905691	PCDH15	5.83E-07	1.18E-03
	14:101142890	43.2491	rs730570	C14orf70	8.32E-07	1.68E-03
	8:73848139	63.9317	rs3735829	KCNB2	1.36E-06	2.75E-03
	10:55602540	51.786	rs12263108	PCDH15	1.97E-06	3.98E-03
	10:55884789	62.7399	rs10825242	PCDH15	2.66E-06	5.37E-03
	22:47836412	55.4207	rs2040411	between LOC339685 and FLJ46257	2.93E-06	5.92E-03
	13:76024169	65.6974	rs9593078	TBC1D4	5.34E-06	1.08E-02
	10:55968685	64.7048	rs10825273	PCDH15	7.18E-06	1.45E-02
	8:73845337	32.2472	rs16919329	KCNB2	7.29E-06	1.47E-02
	11:120121323	45.8118	rs11217775	POU2F3	8.56E-06	1.73E-02
	8:72174978	66.5609	rs1445407	EYA1	8.99E-06	1.82E-02
PC 3	16:71186740	134.2694	rs3094868	HYDIN	1.21E-10	2.45E-07
	1:18956404	62.4871	rs766324	-	2.66E-07	5.37E-04
	16:71186834	135.1328	rs3094869	HYDIN	5.10E-07	1.03E-03
	16:71186860	133.7119	rs3114614	HYDIN	1.70E-06	3.43E-03
	4:42065293	64.8672	rs3804191	SLC30A9	2.10E-06	4.24E-03
	1:18956458	62.5129	rs766325	PAX7	3.23E-06	6.52E-03
	4:42089177	87.0812	rs11051	SLC30A9	4.87E-06	9.83E-03
	4:42065269	65.0554	rs3804190	SLC30A9	4.87E-06	9.84E-03

	16:71186888	112.1089	rs3094870	HYDIN	5.04E-06	1.02E-02
	10:34755348	32.3266	rs1978806	PARD3	5.20E-06	1.05E-02
	2:177992107	52.7546	rs6433650	between RPL29P8 and LOC391465	1.02E-05	2.06E-02
	9:939738	39.7288	rs10977650	DMRT1	1.12E-05	2.25E-02
	4:42050768	42.5554	rs10433709	SLC30A9	1.19E-05	2.39E-02
	4:42003671	39.6328	rs1047626	SLC30A9	1.41E-05	2.84E-02
	9:95325631	72.4649	rs7865019	CENPP	1.64E-05	3.30E-02
PC 4	15:26905021	56.0904	rs17738087	GABRB3	5.14E-10	1.04E-06
	2:216299550	44.1605	rs2577290	FN1	3.20E-09	6.47E-06
	2:216299551	44.4758	rs202099802	FN1	7.68E-08	1.55E-04
	2:216277672	53.8801	rs7588830	FN1	2.02E-07	4.08E-04
	8:73848139	63.9317	rs3735829	KCNB2	3.12E-07	6.31E-04
	2:216276856	53.5923	rs17457252	FN1	4.04E-07	8.16E-04
	11:120121323	45.8118	rs11217775	POU2F3	1.25E-06	2.53E-03
	15:37247246	62.524	rs2122497	MEIS2	1.39E-06	2.80E-03
	4:5570221	50.1162	rs12511039	EVC2	1.70E-06	3.43E-03
	22:46835000	99.9133	rs4823810	CELSR1	1.99E-06	4.02E-03
	13:95096013	76.6697	rs1407995	DCT	2.77E-06	5.59E-03
	1:234330571	58.7768	rs7531988	SLC35F3	3.59E-06	7.24E-03
	12:66114478	77.5812	rs10784490	between PCNPP3 and RPSAP52	3.87E-06	7.82E-03
	10:55884789	62.7399	rs10825242	PCDH15	4.05E-06	8.19E-03
	10:55873113	53.5738	rs4281403	PCDH15	4.37E-06	8.82E-03
	11:120169962	41.0203	rs11217806	POU2F3	4.76E-06	9.61E-03
	10:55935850	65.0369	rs4935501	PCDH15	5.78E-06	1.17E-02
	3:71626262	78.7159	rs830599	FOXP1	6.13E-06	1.24E-02
	8:73845337	32.2472	rs16919329	KCNB2	6.31E-06	1.27E-02
	7:55154688	69.5996	rs2302535	EGFR	6.46E-06	1.31E-02
	14:101142890	43.2491	rs730570	C14orf70	6.74E-06	1.36E-02
	10:55996761	44.9871	rs721825	PCDH15	7.71E-06	1.56E-02
	18:67672106	58.8616	rs56293475	RTTN	7.77E-06	1.57E-02
	18:67741309	38.3198	rs1369293	RTTN	1.06E-05	2.14E-02
PC 8	15:48426484	47.4004	rs1426654	SLC24A5	2.62E-12	5.30E-09
	15:26905021	56.0904	rs17738087	GABRB3	1.34E-10	2.71E-07
	13:34864240	64.5738	rs2065982	between RFC3 and GAMTP2	2.50E-10	5.05E-07
	2:152814028	45.7177	rs16830498	CACNB4	3.09E-09	6.24E-06
	22:46835000	99.9133	rs4823810	CELSR1	4.76E-09	9.61E-06
	12:66114478	77.5812	rs10784490	between PCNPP3 and RPSAP52	7.53E-09	1.52E-05
	2:240282208	86.7675	rs12471054	HDAC4	1.55E-08	3.14E-05
	8:73848139	63.9317	rs3735829	KCNB2	2.81E-08	5.67E-05
	1:234330571	58.7768	rs7531988	SLC35F3	2.85E-08	5.76E-05

	2:216299550	44.1605	rs2577290	FN1	3.44E-08	6.94E-05
	1:36437991	73.6494	rs12096413	EIF2C3	5.89E-08	1.19E-04
	2:216299551	44.4758	rs202099802	FN1	8.57E-08	1.73E-04
	11:120169962	41.0203	rs11217806	POU2F3	1.77E-07	3.57E-04
	14:52607967	47.6033	rs2357442	between COX5AP2 and PTGDR	1.80E-07	3.64E-04
	13:102824147	43.845	rs9557826	FGF14	2.26E-07	4.56E-04
	11:120170030	43.5	rs11217807	POU2F3	2.50E-07	5.06E-04
	2:16815844	42.0387	rs16982612	FAM49A	4.46E-07	9.01E-04
	1:75609049	37.4354	rs12041465	LHX8	4.76E-07	9.61E-04
	2:216276856	53.5923	rs17457252	FN1	4.93E-07	9.96E-04
	13:102816760	59.7048	rs2607642	FGF14	6.37E-07	1.29E-03
	2:149234861	80.7509	rs7569399	MBD5	6.50E-07	1.31E-03
	1:79225423	84.8023	rs1937025	-	7.57E-07	1.53E-03
	2:240295613	59.3506	rs2176046	HDAC4	8.00E-07	1.62E-03
	6:7137363	41.5904	rs1285879	RREB1	9.08E-07	1.84E-03
	X:39934488	75.9539	rs6610384	BCOR	1.04E-06	2.11E-03
	13:95096013	76.6697	rs1407995	DCT	1.09E-06	2.20E-03
PC 9	2:109556761	63.1882	rs260710	EDAR	9.24E-23	1.87E-19
	2:109579738	62.7011	rs260690	EDAR	1.57E-21	3.18E-18
	2:109557099	54.6218	rs260711	EDAR	2.77E-19	5.60E-16
	2:109556667	66.5683	rs260709	EDAR	5.77E-19	1.17E-15
	1:36426398	28.9502	rs7521611	EIF2C3	3.37E-18	6.80E-15
	1:36424338	43.6421	rs716924	EIF2C3	1.18E-17	2.38E-14
	15:48426484	47.4004	rs1426654	SLC24A5	2.15E-17	4.33E-14
	2:109556365	49.2528	rs6542787	EDAR	5.53E-17	1.12E-13
	2:109556418	48.8413	rs7598206	EDAR	7.36E-17	1.49E-13
	2:109567925	64.8321	rs260700	EDAR	1.54E-16	3.10E-13
	2:109562495	67.6753	rs260714	EDAR	3.12E-16	6.30E-13
	17:19175317	65.7768	rs28591622	EPN2	5.42E-16	1.09E-12
	1:36099200	80.6328	rs6668101	PSMB2	2.63E-15	5.30E-12
	1:36437991	73.6494	rs12096413	EIF2C3	3.28E-15	6.62E-12
	17:19239432	42.9114	rs1043809	EPN2	1.71E-14	3.45E-11
	13:102778223	39.8653	rs66578550	FGF14	2.35E-14	4.74E-11
	17:19204863	56.3708	rs4924980	EPN2	3.08E-14	6.23E-11
	3:58091098	50.4041	rs7638552	FLNB	6.43E-14	1.30E-10
	13:102778196	39.5387	rs16959781	FGF14	6.80E-14	1.37E-10
	17:19224397	62.5221	rs6587216	EPN2	7.57E-14	1.53E-10
	10:55949899	34.7048	rs16905691	PCDH15	7.82E-14	1.58E-10
	12:66115201	52.0996	rs12300373	between PCNPP3 and RPSAP52	8.29E-14	1.67E-10
	3:58112556	54.9797	rs2362905	FLNB	1.08E-13	2.18E-10
	17:19247075	43.8229	rs4924987	B9D1	1.09E-13	2.20E-10
	11:120169962	41.0203	rs11217806	POU2F3	1.19E-13	2.41E-10

	10:55971427	55.2399	rs11004141	PCDH15	1.47E-13	2.96E-10
	17:19172505	37.4594	rs28760541	EPN2	1.94E-13	3.92E-10
	17:19211073	40.4834	rs8072587	EPN2	1.94E-13	3.93E-10
	1:234330571	58.7768	rs7531988	SLC35F3	2.46E-13	4.96E-10
	2:109616376	35.0627	rs17034770	near EDAR		5.74E-10
PC 11	22:46835000	99.9133	rs4823810	CELSR1	3.96E-29	8.00E-26
	1:234330571	58.7768	rs7531988	SLC35F3	1.67E-28	3.37E-25
	2:109556761	63.1882	rs260710	EDAR	3.52E-26	7.11E-23
	1:36437991	73.6494	rs12096413	EIF2C3	5.84E-26	1.18E-22
	2:109556667	66.5683	rs260709	EDAR	5.32E-25	1.07E-21
	11:120169962	41.0203	rs11217806	POU2F3	1.11E-23	2.25E-20
	11:120170030	43.5	rs11217807	POU2F3	5.14E-23	1.04E-19
	1:36426398	28.9502	rs7521611	EIF2C3	8.34E-23	1.68E-19
	15:48426484	47.4004	rs1426654	SLC24A5	1.85E-22	3.74E-19
	2:109579738	62.7011	rs260690	EDAR	8.37E-22	1.69E-18
	12:66114478	77.5812	rs10784490	-	1.37E-21	2.76E-18
	2:109557099	54.6218	rs260711	EDAR	1.53E-21	3.10E-18
	13:102824147	43.845	rs9557826	FGF14	1.69E-21	3.41E-18
	15:26905021	56.0904	rs17738087	GABRB3	4.53E-21	9.16E-18
	2:109567925	64.8321	rs260700	EDAR	4.94E-21	9.99E-18
	2:109616376	35.0627	rs17034770	near EDAR	6.53E-21	1.32E-17
	1:36099200	80.6328	rs6668101	PSMB2	1.27E-20	2.57E-17
	11:120121323	45.8118	rs11217775	POU2F3	1.42E-19	2.87E-16
	8:73848139	63.9317	rs3735829	KCNB2	1.55E-19	3.13E-16
	2:109586371	63.5959	rs11123719	EDAR	2.94E-19	5.94E-16
	11:120129533	55.4446	rs11217777	POU2F3	3.18E-19	6.43E-16
	9:125799927	43.821	rs10985869	RABGAP1	3.96E-19	7.99E-16
	13:34864240	64.5738	rs2065982	between RFC3 and GAMTP2	2.53E-18	5.11E-15
	13:102821327	44.81	rs4771420	FGF14	2.88E-18	5.83E-15
	17:19224397	62.5221	rs6587216	EPN2	2.94E-18	5.95E-15
	10:55949151	88.9483	rs16905686	PCDH15	5.29E-18	1.07E-14
	10:55935850	65.0369	rs4935501	PCDH15	5.94E-18	1.20E-14
	14:97277005	61.9908	rs722869	VRK1	9.44E-18	1.91E-14
	11:120126277	40.2825	rs2715881	POU2F3	1.05E-17	2.11E-14
	1:36424338	43.6421	rs716924	EIF2C3	1.26E-17	2.55E-14
PC 12	17:19172505	37.4594	rs28760541	EPN2	2.12E-29	4.28E-26
	12:66115201	52.0996	rs12300373	between PCNPP3 and RPSAP52	2.67E-28	5.38E-25
	13:102816760	59.7048	rs2607642	FGF14	3.10E-28	6.25E-25
	3:58118555	85.8007	rs12632456	FLNB	2.20E-27	4.45E-24
	17:19175317	65.7768	rs28591622	EPN2	2.98E-27	6.02E-24
	17:19211073	40.4834	rs8072587	EPN2	3.36E-27	6.78E-24
	17:19204863	56.3708	rs4924980	EPN2	3.94E-27	7.96E-24

	3:58114655	48.4779	rs9311669	FLNB	1.07E-26	2.17E-23
	11:120118498	74.0258	rs2847502	POU2F3	4.05E-26	8.18E-23
	3:58006600	58.8173	rs839241	FLNB	3.30E-25	6.67E-22
	17:19239432	42.9114	rs1043809	EPN2	3.64E-25	7.36E-22
	11:120130512	46.8339	rs1941411	POU2F3	4.10E-25	8.29E-22
	3:58112556	54.9797	rs2362905	FLNB	4.17E-25	8.41E-22
	1:234330571	58.7768	rs7531988	SLC35F3	6.03E-25	1.22E-21
	3:58112488	57.8506	rs2362904	FLNB	6.43E-25	1.30E-21
	11:120133494	98.3911	rs2715883	POU2F3	1.30E-24	2.63E-21
	3:58113930	49.703	rs2362907	FLNB	1.48E-24	2.98E-21
	1:36437991	73.6494	rs12096413	EIF2C3	3.85E-24	7.78E-21
	17:19247075	43.8229	rs4924987	B9D1	4.30E-24	8.68E-21
	12:112843363	49.4852	rs2301723	RPL6	1.11E-23	2.24E-20
	3:58091098	50.4041	rs7638552	FLNB	2.96E-23	5.97E-20
	8:73848139	63.9317	rs3735829	KCNB2	4.25E-23	8.58E-20
	11:120107411	46.9889	rs882856	POU2F3	4.44E-23	8.97E-20
	14:101142890	43.2491	rs730570	C14orf70	1.19E-22	2.40E-19
	2:109557099	54.6218	rs260711	EDAR	1.56E-22	3.14E-19
	15:26905021	56.0904	rs17738087	GABRB3	2.87E-22	5.80E-19
	6:145055331	52.5387	rs4463276	UTRN	3.10E-22	6.26E-19
	2:109562495	67.6753	rs260714	EDAR	3.26E-22	6.59E-19
	13:34864240	64.5738	rs2065982	between RFC3 and GAMTP2	4.73E-22	9.55E-19
	1:36367780	89.7804	rs595961	EIF2C1	5.83E-22	1.18E-18
PC 13	17:19239432	42.9114	rs1043809	EPN2	7.53E-11	1.52E-07
	17:19211073	40.4834	rs8072587	EPN2	1.25E-10	2.51E-07
	17:19172505	37.4594	rs28760541	EPN2	2.13E-10	4.29E-07
	16:71186860	133.7119	rs3114614	HYDIN	2.49E-10	5.02E-07
	17:19175317	65.7768	rs28591622	EPN2	2.61E-10	5.28E-07
	13:102778223	39.8653	rs66578550	FGF14	3.03E-10	6.11E-07
	17:19204863	56.3708	rs4924980	EPN2	1.89E-09	3.81E-06
	17:19247075	43.8229	rs4924987	B9D1	1.99E-09	4.02E-06
	13:102778196	39.5387	rs16959781	FGF14	3.10E-09	6.27E-06
	17:19224397	62.5221	rs6587216	EPN2	7.38E-09	1.49E-05
	11:120130512	46.8339	rs1941411	POU2F3	1.32E-08	2.67E-05
	2:109579738	62.7011	rs260690	EDAR	1.36E-08	2.75E-05
	11:120133494	98.3911	rs2715883	POU2F3	1.61E-08	3.25E-05
	1:36133918	52.0491	rs677661	between PSMB2 and C1orf216	9.52E-08	1.92E-04
	2:109557099	54.6218	rs260711	EDAR	1.15E-07	2.31E-04
	17:19265797	20.9685	rs10445411	B9D1	1.38E-07	2.78E-04
	1:36426398	28.9502	rs7521611	EIF2C3	1.53E-07	3.08E-04
	2:109562495	67.6753	rs260714	EDAR	1.53E-07	3.10E-04
	3:58112488	57.8506	rs2362904	FLNB	1.64E-07	3.31E-04
	2:109556761	63.1882	rs260710	EDAR	1.64E-07	3.31E-04
	2:109557036	51.8727	rs11123718	EDAR	1.77E-07	3.57E-04
	12:112843363	49.4852	rs2301723	RPL6	1.81E-07	3.65E-04

PC 14	2:109579738	62.7011	rs260690	EDAR	1.01E-20	2.03E-17
	2:109556761	63.1882	rs260710	EDAR	2.99E-20	6.03E-17
	2:109557099	54.6218	rs260711	EDAR	7.56E-18	1.53E-14
	1:36424338	43.6421	rs716924	EIF2C3	2.08E-17	4.20E-14
	2:109556667	66.5683	rs260709	EDAR	3.14E-17	6.35E-14
	17:19175317	65.7768	rs28591622	EPN2	3.89E-17	7.86E-14
	1:36426398	28.9502	rs7521611	EIF2C3	1.22E-16	2.47E-13
	2:109567925	64.8321	rs260700	EDAR	3.69E-16	7.46E-13
	2:109562495	67.6753	rs260714	EDAR	5.90E-16	1.19E-12
	12:66115201	52.0996	rs12300373	between PCNPP3 and RPSAP52	6.95E-16	1.40E-12
	17:19239432	42.9114	rs1043809	EPN2	3.27E-15	6.60E-12
	2:109556365	49.2528	rs6542787	EDAR	3.50E-15	7.06E-12
	17:19204863	56.3708	rs4924980	EPN2	4.52E-15	9.13E-12
	2:109556418	48.8413	rs7598206	EDAR	7.81E-15	1.58E-11
	1:36099200	80.6328	rs6668101	PSMB2	1.41E-14	2.85E-11
	13:102778223	39.8653	rs66578550	FGF14	1.63E-14	3.28E-11
	3:58091098	50.4041	rs7638552	FLNB	2.02E-14	4.08E-11
	1:36437991	73.6494	rs12096413	EIF2C3	2.02E-14	4.08E-11
	3:58112556	54.9797	rs2362905	FLNB	2.26E-14	4.57E-11
	17:19247075	43.8229	rs4924987	B9D1	3.07E-14	6.20E-11
	13:102778196	39.5387	rs16959781	FGF14	3.36E-14	6.79E-11
	3:58112488	57.8506	rs2362904	FLNB	4.35E-14	8.78E-11
	17:19172505	37.4594	rs28760541	EPN2	6.72E-14	1.36E-10
	11:120169962	41.0203	rs11217806	POU2F3	6.85E-14	1.38E-10
	17:19211073	40.4834	rs8072587	EPN2	7.32E-14	1.48E-10
	3:58114655	48.4779	rs9311669	FLNB	1.13E-13	2.27E-10
	17:19224397	62.5221	rs6587216	EPN2	1.22E-13	2.46E-10
	10:55949899	34.7048	rs16905691	PCDH15	1.24E-13	2.50E-10
	3:58118555	85.8007	rs12632456	FLNB	1.49E-13	3.02E-10
	13:102821327	44.81	rs4771420	FGF14	1.67E-13	3.37E-10
PC 15	1:234330571	58.7768	rs7531988	SLC35F3	8.24E-50	1.66E-46
	22:46835000	99.9133	rs4823810	CELSR1	5.65E-46	1.14E-42
	1:36437991	73.6494	rs12096413	EIF2C3	1.10E-45	2.23E-42
	11:120169962	41.0203	rs11217806	POU2F3	1.02E-44	2.07E-41
	8:73848139	63.9317	rs3735829	KCNB2	1.35E-42	2.72E-39
	15:26905021	56.0904	rs17738087	GABRB3	2.79E-42	5.63E-39
	13:102824147	43.845	rs9557826	FGF14	3.24E-41	6.54E-38
	11:120170030	43.5	rs11217807	POU2F3	7.63E-41	1.54E-37
	1:36426398	28.9502	rs7521611	EIF2C3	3.63E-40	7.33E-37
	1:36099200	80.6328	rs6668101	PSMB2	1.69E-37	3.41E-34
	13:102816760	59.7048	rs2607642	FGF14	2.22E-37	4.49E-34
	2:109579738	62.7011	rs260690	EDAR	5.99E-37	1.21E-33
	12:66114478	77.5812	rs10784490	between PCNPP3 and RPSAP52	1.20E-36	2.43E-33

	11:120121323	45.8118	rs11217775	POU2F3	2.49E-36	5.03E-33
	2:109556667	66.5683	rs260709	EDAR	4.41E-36	8.90E-33
	10:55935850	65.0369	rs4935501	PCDH15	3.71E-35	7.49E-32
	2:109567925	64.8321	rs260700	EDAR	7.64E-35	1.54E-31
	11:120129533	55.4446	rs11217777	POU2F3	1.82E-34	3.68E-31
	13:102821327	44.81	rs4771420	FGF14	1.85E-34	3.74E-31
	1:36424338	43.6421	rs716924	EIF2C3	3.71E-34	7.49E-31
	8:73845337	32.2472	rs16919329	KCNB2	1.14E-33	2.31E-30
	2:109556761	63.1882	rs260710	EDAR	1.16E-33	2.35E-30
	13:34864240	64.5738	rs2065982	between RFC3 and GAMTP2	1.94E-33	3.91E-30
	10:55949151	88.9483	rs16905686	PCDH15	2.72E-33	5.50E-30
	2:109557099	54.6218	rs260711	EDAR	5.14E-33	1.04E-29
	3:58118555	85.8007	rs12632456	FLNB	1.73E-32	3.49E-29
	2:152829657	91.5424	rs6721518	CACNB4	3.26E-32	6.57E-29
	12:66115201	52.0996	rs12300373	between PCNPP3 and RPSAP52	6.21E-32	1.25E-28
	2:240282208	86.7675	rs12471054	HDAC4	2.02E-31	4.08E-28
	3:58091098	50.4041	rs7638552	FLNB	2.35E-31	4.75E-28
PC 16	17:19172505	37.4594	rs28760541	EPN2	3.19E-24	6.45E-21
	17:19175317	65.7768	rs28591622	EPN2	6.40E-24	1.29E-20
	17:19204863	56.3708	rs4924980	EPN2	7.78E-24	1.57E-20
	18:19765739	45.5203	rs16964572	GATA6	1.14E-22	2.30E-19
	11:120133494	98.3911	rs2715883	POU2F3	1.97E-22	3.97E-19
	17:19239432	42.9114	rs1043809	EPN2	7.72E-22	1.56E-18
	17:19211073	40.4834	rs8072587	EPN2	1.86E-21	3.75E-18
	8:116776330	42.4465	rs10505268	TRPS1	1.89E-21	3.83E-18
	12:66168151	69.1679	rs7134682	-	5.90E-21	1.19E-17
	8:116700847	40.9004	rs7842702	TRPS1	6.50E-21	1.31E-17
	11:120130512	46.8339	rs1941411	POU2F3	8.60E-21	1.74E-17
	10:34755348	32.3266	rs1978806	PARD3	1.14E-20	2.30E-17
	12:66115201	52.0996	rs12300373	between PCNPP3 and RPSAP52	2.98E-20	6.02E-17
	6:145055331	52.5387	rs4463276	UTRN	6.70E-20	1.35E-16
	17:19247075	43.8229	rs4924987	B9D1	7.21E-20	1.46E-16
	13:102816760	59.7048	rs2607642	FGF14	8.02E-19	1.62E-15
	8:116673229	40.3579	rs10955754	TRPS1	1.10E-18	2.23E-15
	8:116670666	58.3266	rs10505261	TRPS1	1.86E-18	3.76E-15
	11:66910826	78.0111	rs4375446	KDM2A	2.73E-18	5.50E-15
	1:36367780	89.7804	rs595961	EIF2C1	3.87E-18	7.82E-15
	12:66256395	58.5129	rs11175944	HMGA2	6.78E-18	1.37E-14
	7:107704688	66.2269	rs10257477	LAMB4	1.14E-17	2.31E-14
	11:66889299	44.5055	rs4244819	KDM2A	1.31E-17	2.64E-14
	8:116823046	51.3579	rs6981915	near EIF3H	3.47E-17	7.01E-14
	X:39944072	44.6808	rs5963734	BCOR	3.80E-17	7.67E-14
	11:64532579	68.7435	rs523200	SF1	3.81E-17	7.69E-14
	11:120118498	74.0258	rs2847502	POU2F3	4.44E-17	8.97E-14

	14:101142890	43.2491	rs730570	C14orf70	7.78E-17	1.57E-13
	X:39934488	75.9539	rs6610384	BCOR	1.22E-16	2.46E-13
	8:72127562	34.5258	rs79867447	EYA1	2.08E-16	4.20E-13
PC 17	13:102824147	43.845	rs9557826	FGF14	1.08E-37	2.19E-34
	1:36099200	80.6328	rs6668101	PSMB2	9.32E-37	1.88E-33
	13:102816760	59.7048	rs2607642	FGF14	3.53E-36	7.12E-33
	1:36437991	73.6494	rs12096413	EIF2C3	4.84E-36	9.78E-33
	1:36426398	28.9502	rs7521611	EIF2C3	5.28E-34	1.07E-30
	11:120169962	41.0203	rs11217806	POU2F3	7.27E-34	1.47E-30
	2:109579738	62.7011	rs260690	EDAR	1.39E-33	2.81E-30
	13:102821327	44.81	rs4771420	FGF14	1.87E-33	3.77E-30
	11:120170030	43.5	rs11217807	POU2F3	2.86E-33	5.78E-30
	17:19224397	62.5221	rs6587216	EPN2	6.78E-33	1.37E-29
	17:19175317	65.7768	rs28591622	EPN2	4.91E-32	9.91E-29
	17:19172505	37.4594	rs28760541	EPN2	5.31E-32	1.07E-28
	17:19204863	56.3708	rs4924980	EPN2	8.39E-32	1.69E-28
	22:46835000	99.9133	rs4823810	CELSR1	3.26E-31	6.59E-28
	8:73848139	63.9317	rs3735829	KCNB2	4.08E-31	8.24E-28
	1:234330571	58.7768	rs7531988	SLC35F3	6.60E-31	1.33E-27
	2:240282208	86.7675	rs12471054	HDAC4	1.03E-30	2.09E-27
	13:102778223	39.8653	rs66578550	FGF14	1.09E-30	2.21E-27
	17:19211073	40.4834	rs8072587	EPN2	1.87E-30	3.77E-27
	13:102778196	39.5387	rs16959781	FGF14	5.24E-30	1.06E-26
	17:19239432	42.9114	rs1043809	EPN2	7.00E-30	1.41E-26
	2:109567925	64.8321	rs260700	EDAR	9.50E-30	1.92E-26
	13:34864240	64.5738	rs2065982	between RFC3 and GAMTP2	1.95E-29	3.95E-26
	12:112843363	49.4852	rs2301723	RPL6	9.99E-29	2.02E-25
	2:109556667	66.5683	rs260709	EDAR	5.24E-28	1.06E-24
	11:120118498	74.0258	rs2847502	POU2F3	1.28E-27	2.58E-24
	3:58118555	85.8007	rs12632456	FLNB	1.47E-27	2.98E-24
	2:109562495	67.6753	rs260714	EDAR	1.50E-27	3.03E-24
	8:73845337	32.2472	rs16919329	KCNB2	2.44E-27	4.92E-24
	1:36424338	43.6421	rs716924	EIF2C3	2.49E-27	5.03E-24
PC 19	2:109556761	63.1882	rs260710	EDAR	5.93E-18	1.20E-14
	2:109579738	62.7011	rs260690	EDAR	5.46E-17	1.10E-13
	2:109557099	54.6218	rs260711	EDAR	7.39E-16	1.49E-12
	10:55971427	55.2399	rs11004141	PCDH15	3.79E-15	7.66E-12
	2:109556667	66.5683	rs260709	EDAR	1.02E-14	2.07E-11
	3:58091098	50.4041	rs7638552	FLNB	1.38E-14	2.78E-11
	2:109567925	64.8321	rs260700	EDAR	3.45E-14	6.97E-11
	3:58112556	54.9797	rs2362905	FLNB	3.92E-14	7.92E-11
	12:66115201	52.0996	rs12300373	between PCNPP3 and RPSAP52	7.29E-14	1.47E-10
	3:58118555	85.8007	rs12632456	FLNB	7.48E-14	1.51E-10
	3:58114655	48.4779	rs9311669	FLNB	7.59E-14	1.53E-10

	3:58112488	57.8506	rs2362904	FLNB	7.71E-14	1.56E-10
	10:55968685	64.7048	rs10825273	PCDH15	1.34E-13	2.70E-10
	1:36426398	28.9502	rs7521611	EIF2C3	1.72E-13	3.47E-10
	11:120169962	41.0203	rs11217806	POU2F3	1.80E-13	3.63E-10
	2:109562495	67.6753	rs260714	EDAR	2.08E-13	4.20E-10
	1:36437991	73.6494	rs12096413	EIF2C3	2.76E-13	5.58E-10
	1:36424338	43.6421	rs716924	EIF2C3	3.02E-13	6.09E-10
	1:234330571	58.7768	rs7531988	SLC35F3	4.01E-13	8.10E-10
	10:55949151	88.9483	rs16905686	PCDH15	4.09E-13	8.25E-10
	3:58113930	49.703	rs2362907	FLNB	4.92E-13	9.93E-10
	10:55996761	44.9871	rs721825	PCDH15	4.97E-13	1.00E-09
	3:58144340	62.6845	rs6787425	FLNB	8.64E-13	1.74E-09
	3:58109162	64.0148	rs1131356	FLNB	1.44E-12	2.92E-09
	11:120170030	43.5	rs11217807	POU2F3	1.48E-12	2.98E-09
	17:19175317	65.7768	rs28591622	EPN2	2.11E-12	4.26E-09
	10:55935850	65.0369	rs4935501	PCDH15	2.46E-12	4.96E-09
	10:55949899	34.7048	rs16905691	PCDH15	3.68E-12	7.43E-09
	14:97277005	61.9908	rs722869	VRK1	4.85E-12	9.79E-09
	11:120129533	55.4446	rs11217777	POU2F3	1.18E-11	2.38E-08

In general, all the principal components demonstrated higher association values (lower p-values) than the actual craniofacial measurements they were representing. This is not surprising, given the fact that each principal component represents numerous measurements and as a result may demonstrate association with several markers. A fact that is even more important, is that principal components may represent the craniofacial "vectors" (or shapes), which in essence indicate the morphological direction (or areas) in which a group of cells is migrated during proliferation and formation of specific facial structures. This complex process is clearly regulated by hundreds of genes, while a few of them were specified in the current project.

Given a large number of significantly associated markers and genes, the following section briefly outlines only a few, which demonstrated the most significant association with the principal components.

Principal component number 15 demonstrated the most significant p-values and was associated with 30 markers in 13 genes and 2 intergenic regions, such as SLC35F3, CELSR1, EIF2C3, POU2F3, KCNB2, GABRB3, FGF14, PSMB2 genes. The majority of these genes previously were not associated with craniofacial morphology or related

genetic syndromes. The most significant association was demonstrated with intronic SNP rs7531988 (p-value 8.24E-50), located in the Solute Carrier Family 35 Member F3 (SLC35F3) gene, which encodes solute transporter, although its specific function is unknown. The same gene was also associated with the nasal vertical prominence angle (Table 51). Notably, an important paralog of this gene is SLC35F4, which was linked to bipolar disorder as a part of a GWAS study [468]. Bipolar disorder is a serious mental illness, related to schizophrenia. Interestingly, schizophrenia was previously associated with distinctive facial features, as discussed in Chapter 1 and elsewhere [109, 110]. Another protein of the same Solute Carrier Family 35, SLC35D1 was previously linked to the Schneckenbecken Dysplasia condition (OMIM:269250), which is associated with several symptoms, such as macrocephaly, depressed midface, superliorly oriented orbits and short neck.

Another important gene, significantly associated with multiple principal components is EIF2C3. This gene encodes a protein that has a similar function to EIF2C1, which has been previously described in this section. KCNB2 gene encodes a potassium voltage-gated channel required for epithelial electrolyte transport. CELSR1, POU2F3 and GABRB3 genes were found associated with specific craniofacial measurements and briefly described previously in this section. On the contrary, the fibroblast growth factor (FGF14) was shown to be involved in a variety of biological processes, including embryonic development, cell growth and morphogenesis

SNP annotation analysis using SNPnexus revealed that of the 161 significant markers, the majority were found in introns. While only twelve (12) are located in coding regions, of which 6 are non-synonymous SNPs. Functional prediction by both SIFT and PolyPhen demonstrated however, that these amino acid substitutions would probably have no significant effect on the protein function. Ten more SNPs (out of 161 markers) are located in either 3' or 5' genomic regions, which are known to be involved in gene regulation. Five of these SNPs (rs61726477, rs1937025, rs16905686, rs2715881 and rs56293475) were predicted to affect transcription binding sites. Regulome analysis of all significant markers predicted 126 markers to have a regulatory function. This finding can be seen as a confirmation for the 2013 publication (discussed previously in this section), which utilized mouse model and found that transcriptional enhancers may regulate fine-tuning of the craniofacial shape and contribute to the spectrum of facial variation found in human population [213].

In conclusion, this study has identified numerous genetic markers in a set of candidate genes, significantly associated with various visible craniofacial traits. All the

associations revealed in this study are novel and have not been previously reported. The results of the association analysis also demonstrated a complex level of interactions between the genes involved in the craniofacial embryonic development. Despite the importance of the association between specific 3D measurements and SNPs, the association of principal components provide more composite information. Given that embryonic developmental processes such as cells proliferation, polarity orientation and migration occur in the 3D environment, principal components that in essence represent specific facial shapes (or "vectors") provide a more accurate representation of these processes. It should be emphasized though, that given the nature of the craniofacial traits, these complex traits are regulated by many different genes. This aspect highlights the demand for additional association studies, incorporating dense SNP panels and even more importantly, better coverage of the craniofacial characteristics, perhaps with a dense polygon mesh.

The identification of associated genetic markers in this project forms a foundation for designing statistical models and subsequent assays for prediction of facial appearance. However, keeping in mind the recent human height studies association result, which screened more than 180,000 individuals and found more than 180 significant variants; the prediction of the facial appearance would probably be a more complicated task [298-300, 469].

5.4. Assessment of statistical models for prediction of EVTs and ancestry

5.4.1. Introduction

A comprehensive EVT prediction assay is expected to include both AIMs and pigmentation-informative markers, providing extensive information for investigative purposes. This approach can be called a "meta-analysis" as it will analyse and predict the likelihood of specific external appearance, most likely by classifying an unknown individual with a relevant database hit. Although it is not likely that predictive kit in the near future will produce an image comparable to a photograph, this tool can provide an independent means for validating eyewitness testimony, (given the notorious inaccuracy of eyewitness description) or supply valuable information in the absence of such.

This section summarizes the results of a preliminary statistical analysis for prediction of several pigmentation and craniofacial traits and ancestry. Due to a time limit of this project, only a small subset of traits was evaluated for prediction purposes, and the results presented in this section are currently undergoing further analysis. Regardless, these results demonstrate a first attempt to create statistical models for prediction of the craniofacial traits, which has not been performed previously.

5.4.2. Methods

Statistical modelling was performed using the same genotype and phenotype data, which were detailed in Section 5.2. Prediction model was based on the forward stepwise conditional logistic regression, using PASW Statistics 18 (SPSS, Inc., 2009, Chicago, IL). In this model, each marker is added at a time and after each step, the model measures a pseudo-R^2 (Nagelkerke) value of the output to assess its predictive power. Following this algorithm, each step of the regression has iterated over all alleles and included them in the model when the p-value for prediction of the phenotype was improved.

The PASW Statistics algorithm used for the logistic regression analysis removed samples, if even one datum point (genotype or phenotype) used to generate the predictive model was missing. Given that algorithm was not tolerant for missing data, an imputation approach using BEAGLE software was undertaken. This software package applies a localised haplotype-cluster model by alignment of missing genotypes

in the analysed sample set to other individuals in the relevant population [470]. The imputation resulted in 1,642 fully genotyped markers in 567 individuals. As a result, the PASW Statistics was able to predict phenotypes of these samples.

For prediction purposes, each of the numeric craniofacial distances was artificially divided into five bins, by finding the minimum and maximum value of each phenotype and calculating their difference. Subsequently, a calculation of 0.2 x 'difference' was used as a "step size" function, to obtain five bins per each trait. To test the accuracy of the model, twenty five "blind" samples were randomly removed from the dataset and used for prediction. Based on the prediction analysis, the following statistics were calculated: true and false positives, true and false negatives, sensitivity, specificity, positive predictive value, negative predictive value, positive likelihood ratio (LR+) and negative likelihood ratio (LR-).

5.4.3. Results and discussion

Statistical models were generated for prediction of pigmentation traits, such as blue eyes, brown eyes, brown hair, black hair, fair skin, and black skin; European and Asian ancestries; craniofacial measurements (divided into five bins each), such as cephalic index (CI), zy-zy, n-prn, n-sn and al-al. Each model estimated the phenotype of 567 individuals by analysing their regression with 1,462 genetic markers.

The results of the statistical modelling are summarized in Table 54.

Table 54. Prediction model results for investigated phenotypes (blue eyes, brown eyes, brown hair, black hair, fair skin, and black skin; European and Asian ancestries; cephalic index (CI), zy-zy, n-prn, n-sn and al-al) in 567 individuals. The numbers (1-5) near each craniofacial distance represent artificial bins. Legends: (TP and FP) are true and false positives, (TN and FN) are true and false negatives, (LR+) is a positive likelihood ratio and (LR-) is a negative likelihood ratio.

	TP	TN	FP	FN	Correct	Incorrect	Total	% of correct predictions	Nagelkerke R Square	Sensitivity	Specificity	Positive predictive value	Negative predictive value	LR+	LR-
Brown hair	262	261	19	25	523	44	567	92.2	.850	91.29%	33.29%	93.24%	91.26%	1.4	0.3
Black hair	134	389	19	25	523	44	567	92.2	.855	84.28%	42.65%	87.58%	93.96%	1.5	0.4
Blue eyes	132	372	35	28	504	63	567	88.9	.733	82.50%	91.40%	79.04%	93.00%	9.6	0.2
Brown eyes	274	279	6	8	553	14	567	97.5	.918	97.16%	97.89%	97.86%	97.21%	46.2	0.0
European ancestry	351	194	13	9	545	22	567	96.1	.902	97.50%	93.72%	96.43%	95.57%	15.5	0.0
Asian ancestry	42	511	5	9	553	14	567	97.5	.833	82.35%	99.03%	89.36%	98.27%	85.0	0.2
Fair skin	115	362	37	53	477	90	567	84.1	.598	68.45%	90.73%	75.66%	87.23%	7.4	0.3
Black skin	11	413	142	1	424	143	567	74.8	.657	91.67%	74.41%	7.19%	99.76%	3.6	0.1
CI 1	16	532	2	17	548	19	567	96.6	.525	48.48%	99.63%	88.89%	96.90%	129.5	0.5
CI 2	308	182	42	35	490	77	567	86.4	.760	89.80%	81.25%	88.00%	83.87%	4.8	0.1
CI 3	103	410	20	34	513	54	567	90.5	.774	75.18%	95.35%	83.74%	92.34%	16.2	0.3
CI 4	0	556	0	11	556	11	567	98.1	.450	-	100.00%	-	98.06%	-	-
CI 5	1	565	1	0	566	1	567	99.8	.667	100.00%	99.82%	50.00%	100.00%	566.0	0.0
zy-zy 1	191	290	36	50	481	86	567	84.8	.705	79.25%	88.96%	84.14%	85.29%	7.2	0.2
zy-zy 2	173	344	22	28	517	50	567	91.2	.830	86.07%	93.99%	88.72%	92.47%	14.3	0.1
zy-zy 3	16	526	5	20	542	25	567	95.6	.577	44.44%	99.06%	76.19%	96.34%	47.2	0.6
zy-zy 4	0	565	0	2	565	2	567	99.6	.395	-	100.00%	-	99.65%	-	-
zy-zy 5	3	532	5	27	535	32	567	94.4	.363	10.00%	99.07%	37.50%	95.17%	10.7	0.9
n-prn 1	135	373	23	36	508	59	567	89.6	.745	78.95%	94.19%	85.44%	91.20%	13.6	0.2
n-prn 2	185	314	34	34	499	68	567	88.0	.788	84.47%	90.23%	84.47%	90.23%	8.6	0.2
n-prn 3	52	475	14	26	527	40	567	92.9	.673	66.67%	97.14%	78.79%	94.81%	23.3	0.3
n-prn 4	2	552	0	13	554	13	567	97.7	.765	13.33%	100.00%	100.00%	97.70%		0.9
n-prn 5	7	534	3	23	541	26	567	95.4	.437	23.33%	99.44%	70.00%	95.87%	41.8	0.8
n-sn 1	134	361	32	40	495	72	567	87.3	.758	77.01%	91.86%	80.72%	90.02%	9.5	0.3
n-sn 2	174	315	41	37	489	78	567	86.2	.734	82.46%	88.48%	80.93%	89.49%	7.2	0.2
n-sn 3	53	478	10	26	531	36	567	93.7	.665	67.09%	97.95%	84.13%	94.84%	32.7	0.3
n-sn 4	2	553	1	11	555	12	567	97.9	.486	15.38%	99.82%	66.67%	98.05%	85.2	0.8
n-sn 5	12	526	4	25	538	29	567	94.9	.457	32.43%	99.25%	75.00%	95.46%	43.0	0.7
al-al 1	189	309	34	35	498	69	567	87.8	.737	84.38%	90.09%	84.75%	89.83%	8.5	0.2
al-al 2	169	327	34	37	496	71	567	87.5	.743	82.04%	90.58%	83.25%	89.84%	8.7	0.2
al-al 3	25	508	11	23	533	34	567	94.0	.624	52.08%	97.88%	69.44%	95.67%	24.6	0.5
al-al 4	3	557	3	4	560	7	567	98.8	.556	42.86%	99.46%	50.00%	99.29%	80.0	0.6
al-al 5	5	534	5	23	539	28	567	95.1	.454	17.86%	99.07%	50.00%	95.87%	19.2	0.8

Blue and brown eyes, as well as European and Asian ancestries demonstrated the highest accuracy of prediction (from 88.9% to 97.5%) in the general prediction model. The accuracy and efficiency of prediction was also assessed by evaluating the following parameters: Nagelkerke R^2 (between 0.733 to 0.918), sensitivity (between 82.4% and 97.5%), specificity (between 91.4% and 99.0%), positive predictive value (between 79.0% and 97.9%), negative predictive value (between 93% to 98.3%), positive likelihood ratio (between 9.6 to 85) and negative likelihood ratio (between 0 to 0.2). A brief interpretation of these results is that these models demonstrated high robustness (R^2 values) and high ability to accurately include or exclude an individual from these phenotypes.

Prediction models for brown and black hair demonstrated high accuracy (92.2%), robustness (R^2 = 0.850) and sensitivity (between 84.3% and 91.3%), although low specificity of exclusion. Prediction of the black and fair skin colour demonstrated a moderate accuracy and efficiency of prediction, although the sample size for black skin was of a limit number, which could affect the output.

Prediction of the craniofacial distances demonstrated relatively high accuracy for several phenotypes. It should be noted however, that due to non-equal (non-normal) distribution of numeric values in each bin, some of them do not represent an adequate phenotype for prediction (e.g. CI 1, 4 and 5). As a result, prediction models for only the following phenotypes were evaluated: CI 2, CI 3, zy-zy 1, zy-zy 2, n-prn 1, n-prn 2, n-prn 3, n-sn 1, n-sn 2, n-sn 3, al-al 1 and al-al 2.

In general, all these phenotypes demonstrated a good prediction accuracy (between 84.8% and 93.7%). The goodness-of-fit for these models varied from 0.665 to 0.830 and the sensitivity and specificity showed values between 66.7% to 98.8% and between 81.3% and 98% respectively. The rest of the statistical values, such as negative and positive predictive value and negative and positive likelihood ratio also confirmed the robustness of these models.

5.4.4. Accuracy of prediction on blind samples

To ascertain the validity of the model predictions for each phenotype, binary logistic regressions were run on 25 randomly chosen “blind” samples. The results of phenotype prediction for these samples are summarized in Table 55.

Table 55. Actual prediction results for 25 blind samples. The numbers (1-5) near each craniofacial distance represent artificial bins. Legends: (TP and FP) are true and false positives, (TN and FN) are true and false negatives, (LR+) is a positive likelihood ratio and (LR-) is a negative likelihood ratio.

	TP	TN	FP	FN	Correct	Incorrect	Total	% of correct predictions	Sensitivity	Specificity	Positive predictive value	Negative predictive value	LR+	LR-
Brown hair	8	6	5	6	14	11	25	56.0	57.1%	30.0%	61.5%	50.0%	0.8	1.4
Black hair	1	16	4	4	17	8	25	68.0	20.0%	48.5%	20.0%	80.0%	0.4	1.7
Blue eyes	8	15	2	0	23	2	25	92.0	100.0%	88.2%	80.0%	100.0%	8.5	0.0
Brown eyes	7	16	1	1	23	2	25	92.0	87.5%	94.1%	87.5%	94.1%	14.9	0.1
European ancestry	14	6	1	4	20	5	25	80.0	77.8%	85.7%	93.3%	60.0%	5.4	0.3
Asian ancestry	1	22	0	2	23	2	25	92.0	33.3%	100.0%	100.0%	91.7%	-	0.7
Fair skin	5	11	5	4	16	9	25	64.0	55.6%	68.8%	50.0%	73.3%	1.8	0.6
Black skin	0	20	5	0	20	5	25	80.0	-	80.0%	-	100.0%	-	-
CI 1	0	25	0	0	25	0	25	100.0	-	100.0%	-	100.0%	-	-
CI 2	7	5	7	6	12	13	25	48.0	53.8%	41.7%	50.0%	45.5%	0.9	1.1
CI 3	3	10	4	8	13	12	25	52.0	27.3%	71.4%	42.9%	55.6%	1.0	1.0
CI 4	0	25	0	0	25	0	25	100.0	-	100.0%	-	100.0%	-	-
CI 5	0	24	1	0	24	1	25	96.0	-	96.0%	-	100.0%	-	-
zy-zy 1	7	4	4	10	11	14	25	44.0	41.2%	50.0%	63.6%	28.6%	0.8	1.2
zy-zy 2	2	13	6	4	15	10	25	60.0	33.3%	68.4%	25.0%	76.5%	1.1	1.0
zy-zy 3	0	24	1	0	24	1	25	96.0	-	96.0%	-	100.0%	-	-
zy-zy 4	0	25	0	0	25	0	25	100.0	-	100.0%	-	100.0%	-	-
zy-zy 5	0	23	1	1	23	2	25	92.0	0.0%	95.8%	-	95.8%	0.0	1.0
n-prn 1	3	10	6	6	13	12	25	52.0	33.3%	62.5%	33.3%	62.5%	0.9	1.1
n-prn 2	1	12	5	7	13	12	25	52.0	12.5%	70.6%	16.7%	63.2%	0.4	1.2
n-prn 3	0	19	3	3	19	6	25	76.0	0.0%	86.4%	-	86.4%	0.0	1.2
n-prn 4	0	23	0	2	23	2	25	92.0	0.0%	100.0%	-	92.0%	-	1.0
n-prn 5	0	23	0	2	23	2	25	92.0	0.0%	100.0%	-	92.0%	-	1.0
n-sn 1	2	9	5	9	11	14	25	44.0	18.2%	64.3%	28.6%	50.0%	0.5	1.3
n-sn 2	2	10	10	3	12	13	25	48.0	40.0%	50.0%	16.7%	76.9%	0.8	1.2
n-sn 3	1	20	0	4	21	4	25	84.0	20.0%	100.0%	100.0%	83.3%	-	0.8
n-sn 4	0	24	0	1	24	1	25	96.0	0.0%	100.0%	-	96.0%	-	1.0
n-sn 5	0	21	2	2	21	4	25	84.0	0.0%	91.3%	-	91.3%	0.0	1.1
al-al 1	4	14	3	4	18	7	25	72.0	50.0%	82.4%	57.1%	77.8%	2.8	0.6
al-al 2	7	9	4	5	16	9	25	64.0	58.3%	69.2%	63.6%	64.3%	1.9	0.6
al-al 3	0	23	2	0	23	2	25	92.0	-	92.0%	-	100.0%	-	-
al-al 4	0	25	0	0	25	0	25	100.0	-	100.0%	-	100.0%	-	-
al-al 5	0	19	2	4	19	6	25	76.0	0.0%	90.5%	-	82.6%	0.0	1.1

Prediction of the blue and brown eyes and the Asian ancestry resulted in 92% accuracy and high sensitivity (except for the Asian ancestry), specificity and efficiency. Compared to the model, these values were slightly lower for the brown eye and Asian ancestry and slightly higher for the blue eye prediction (Table 55). The European ancestry and black skin colour prediction resulted in 80% accuracy (compared to 96% and 74.8% in the model respectively) and lower specificity and sensitivity. The prediction accuracies for black and brown hair as well as for fair skin, were significantly lower than predicted by the model (between 56% and 68% compared to between 84% and 92% respectively).

Published data demonstrate over 90% accuracy for blue and brown eye colour, 78.5% for brown and 87.5% for black hair colour on average [296]. Another recent study demonstrated predictions of eye colour being 85 % correct for brown and 70 % correct for blue eyes, predictions of hair colour 72 % accurate for brown and 58 % accurate for black hair and 97% accurate for European and Asian ancestries [377]. Considering published results, it can be summarized that prediction accuracy of blind samples demonstrated in this study, was in consensus with published data.

Prediction of al-al 1 and al-al 2 bins however, demonstrated low prediction accuracy of 72% and 64%, compared to 87.8% and 87.5% respectively in the model. While the prediction of CI 2, CI 3, zy-zy 1, zy-zy 2, n-prn 1, n-prn 2, n-prn 3, n-sn 1, n-sn 2 and n-sn 3 bins was completely unsuccessful (equal to flipping a coin). These results however, maybe misleading, as only 25 blind samples were used for prediction. In addition, most craniofacial phenotypes were underrepresented or represented unequally in their respective bins due to “unweighted” bins separation.

As noted in Section 5.4.1, these results are only preliminary and are currently under further analysis. This analysis includes usage to alternative statistical software in order to reduce the problem of algorithm intolerance of missing data, calculating adequately weighted bins for craniofacial measurements and using a bigger blind sample set for prediction. Given the high association linkage demonstrated by these traits as well as according to the accuracy demonstrated by prediction model, the final prediction of these phenotypes in a larger blind sample set, holds promise to surpass preliminary output.

Chapter 6

Summary, conclusions and future directions

6.1. Introduction

The overall goal of the current study was to identify SNPs involved in the determination of craniofacial morphology. To confirm the statistical methods used, an analysis of the association of pigmentation traits and ancestry with known previously described genetic markers was also performed. The final aim was to incorporate significantly associated SNPs in prediction models for accurate estimation of externally visible traits (EVT) and ancestry in a blind sample set.

Single nucleotide polymorphisms (SNPs) have started to play a progressively important role in the forensic DNA analysis over the last ten years. While short tandem repeats (STRs) continue to be the main "players" in the forensic identification frontier, other fields such as forensic molecular phenotyping (FMP) are predominated by SNP usage. The main "non-identity" fields of forensic interest are pigmentation and ancestry prediction. Estimation of eye, skin and hair colour as well as ancestry of an unknown DNA sample holds great potential as an investigative lead by reducing a pool of potential suspects as well as assisting with uncovering the identity of unknown human remains and mass disaster victims. However, the genetic factors behind the most important part of human visual appearance – the face, remain poorly understood.

This project attempted to find genetic polymorphisms that influence normal variation in specific craniofacial characteristics. Subsequently, these markers can be incorporated with pigmentation, ancestry and identity-informative SNPs into a comprehensive forensic assay.

6.2. Candidate genes and SNPs selection process

In the last two years, a limited number of attempts have been made to identify SNPs that may contribute to normal variation in craniofacial appearance. To date, several markers, associated with bizygomatic distance and several nasal area features have been identified [76, 77]. While both studies have implemented a genome wide association study (GWAS) approach by screening hundreds of thousands SNPs, the current project strived to achieve the same goal using candidate genes approach. The main advantage of the candidate genes study over GWAS is that the former was focused on genes that were previously associated with craniofacial embryogenetics or various inherited craniofacial syndromes, rather than screening hundreds of thousands of non-specific markers. This approach aimed to increase the chances of finding significant associations

between SNPs and visible traits and also to reduce the number of samples needed for robust association analysis.

An extensive literature and web database search for candidate genes and markers revealed 177 genes and intergenic regions comprising approximately 1,200 SNPs, potentially involved in normal craniofacial appearance variation. This list was subsequently extended by 331 markers, previously associated with normal pigmentation, 208 ancestry informative markers, 91 identity informative SNPs, 46 INDELs, 17 autosomal STRs, 15 Y-chromosome STRs, 37 Y-chromosome SNPs and 57 Mitochondrial SNPs. The addition of non-craniofacial markers was made in order to build a comprehensive forensic "Identikit", able to provide complete forensic molecular portrait of a questionable sample, including identity-informative, ancestry-informative, lineage-informative and appearance-informative markers. It should be noted however, that some markers such as mitochondrial SNPs, Y chromosome STRs and approximately 20% of the other markers were not genotyped due to primer design failure. Other markers, such as autosomal STRs and INDELs were not analysed, due to time limits of this project. Nevertheless, the relevant data generated in this study are available for subsequent analysis in a future project and demonstrate one of the first attempts to incorporate different forensic markers in one assay.

6.3. Assessment of reproducibility and normal distribution of the craniofacial measurements

The reproducibility of facial measurements was assessed in 54 linear distances, 22 ratios between linear distances and 10 angular distances in 13 facial images tested on two occasions, by the same examiner. The assessment of the linear craniofacial measurements reproducibility from 3D images, demonstrated between 6 mm and less than 2 mm median difference (MD), with the majority of measurements (62%), showing the lower variation range (≤ 2mm). The evaluation of facial indexes reproducibility resulted in the highest MD of 8.02 mm, although 68.2% of all values demonstrated significantly lower MD of 3 mm. The reproducibility of the angular distances demonstrated very robust results of between 3.75 and 0.38 degrees between paired measurements.

Poor image quality and general difficulty in finding specific craniofacial landmarks in a few 3D images were the main factors responsible for the variability between

measurements observed. Nevertheless, the current study demonstrated generally higher reproducibility, compared to published studies and provided reliable data for genomic association analysis.

Assessment of normal distribution of the collected data is the first essential step in the quality control process, prior to performing an association (or any other) study with genetic markers. While there is no wide scientific consensus on existence of normality in a real dataset, this test helps to identify extreme outliers, which may represent potential errors in measurements that need to be evaluated and corrected if required.

The Shapiro-Wilk test for normality was calculated for the craniofacial dataset that was segregated by sex and six main population groups. This analysis helped to identify several outliers in measurements, resulting from erroneous assignment of craniofacial landmarks and subsequent miscalculation of the relevant distances. Following correction of the miscalculated values, the majority of the measurements demonstrated normal distribution and allowed a robust and more accurate association study.

The evaluation of such a big variety of craniofacial measurements (n=92) in different population groups provided an opportunity to collect valuable data, considering a very limited availability of similar data in published sources. While many recorded measurements were consistent with the published data, several facial distances produced data, which were unavailable previously in the specific population groups (especially from 3D images), providing novel and useful information for anthropological research and aesthetics medicine application. It should be noted however, that a larger sample set (especially of specific ethnicities) is required to replicate the results of this study and make further conclusions.

Principal component analysis **(**PCA) was applied on the craniofacial measurements in order to reduce the dimensionality of the data by identifying eigenvectors, along which the variation in the data is expected to be maximal. The PCA identified 20 principal components, which accounted for almost 90% of total variance (Table 32 and Figure 52), while the first three and another seven components explained 1/3 and 2/3 of the variance respectively.

The validity of PCA was confirmed by the fact that the craniofacial distances in each principal component were automatically grouped according to their similarity in the

Euclidean space orientation and corresponding landmarks (Figure 52). The main reason behind PCA was to use only these components for subsequent association analysis, rather than all the individual measurements, due to time limits. However, eventually all the individual measurements and all the principal components were analysed for association with genetic markers.

An additional advantage of PCA was that these eigenvectors provided a more adequate representation of the craniofacial morphological structure than individual distances, considering the complexity of facial structures formation during the embryonic development. In fact, the associations of principal components with genetic markers were stronger than of individual measurements, providing valuable information on novel genes involved in the craniofacial morphology and the opportunity to incorporate these markers in prediction models.

6.4. A pilot study for evaluation of the GoldenGate platform for SNP genotyping of pristine and degraded DNA samples

Supplementary to phenotype prediction capabilities, SNPs offer a greater success rate for genotyping severely degraded DNA samples due to shorter regions of amplified genome, often impossible with routine STR typing. This advantage was investigated by genotyping pristine (n=57) and environmentally challenged (n=42) DNA samples of various quantity (20 – 250 ng), implementing both manufacturer recommended and complementary methods of whole genome amplification (WGA).

The sensitivity study demonstrated that the GoldenGate assay is tolerant to a more than fivefold decrease in template input (<50 ng vs 250 ng, as recommended by the manufacturer). However, it also demonstrated that the genomic template must be of high quality, without DNA degradation.

When the DNA amount was reduced to 20 ng or less, a significant decrease in performance as demonstrated by poor call rate and poor allele clustering was observed. The exposure of DNA samples to various environmental insults (heat, fresh water and salt water) significantly reduced genotyping performance.

The GoldenGate assay's input amount of between 50 ng to 250 ng DNA template has no practical forensic application (except reference databasing), as routine forensic DNA

typing typically deals with 0.5-1 ng of template. In attempts to assess the performance of this method with low DNA quantities, samples with a range of 10 ng and 80 ng of high molecular weight template were subjected to WGA protocol with several alterations. However, the results of the WGA reactions suggest that even for the high quality DNA sample, a decent quantity of approximately 50 ng template was needed for robust output. In addition, the WGA method introduced an amplification bias, especially in samples of low quantity and quality.

These results demonstrate that despite its relatively high multiplexing ability (up to 96 markers per reaction), GoldenGate assay does not hold promise for application in the routine forensic DNA casework. This conclusion led to assessment of a Massive Parallel Sequencing (MPS) platform, capable of genotyping thousands of target markers from 10 ng (recommended amount) and down to 100 pg of DNA (with additional amplification). The results of this study were published in Journal of Forensic Investigation, 1:1 (8), 2013 [349].

6.5. A pilot study for evaluation of a 96 SNP panel for prediction of pigmentation and ancestry

The visual traits such as eye, hair and skin colour are important characteristics in forensic investigation. These traits are highly heritable and are known to be regulated by many genes. Extensive research in this field disclosed many specific polymorphisms associated with variation in pigmentation that were subsequently incorporated in a number of predictive assays. Among other studies, this project included a pilot study that investigated performance of the 96-plex SNP assay for prediction of pigmentation and ancestry in 917 DNA samples from three main US population groups (African Americans, Hispanics and Caucasians). Given that certain pigmentation-informative SNPs overlap with AIMs, a relatively small subset of markers can provide enhanced assistance by predicting both pigmentation phenotype and ancestry [12, 471]. This topic was explored in the current project and demonstrated relatively high prediction accuracy for ancestry (84-94% accuracy), although lower accuracy for eye colour (between 71% and 89%) and hair and skin colour (between 47% and 85%). The explanation for relatively low performance of this assay lay in the significant variation in DNA quality and quantity of samples as well as variability between different assay runs, which

resulted in many samples and loci with missing data. Given that the algorithm used for prediction analysis was intolerant for missing data, approximately 20% of samples and 28% of SNPs did not meet the requirements for statistical analysis and were exclided. Altogether, this led to decreased sensitivity of the prediction model.

In conclusion, the 96-plex GoldenGate assay demonstrated a relatively low performance considering its high demand for DNA quantity and quality and consequently a lower-than-expected prediction ability. Nevertheless, this pilot study was an important performance test for potential forensic use of a high multiplex SNP assay and an opportunity to develop predictive models, which were used for subsequent study.

The results of this study were submitted for publication in the Journal of Forensic Investigation.

6.6. Ion Torrent platform evaluation study

The use of Ion Torrent is relatively new in the research field and completely novel in the forensic DNA area. Compared to the GoldenGate assay, it offers almost unlimited multiplex capacity and significantly lower DNA input (recommended as 10 ng, which although can be decreased to 100 pg with additional amplification steps). Both platforms however, are time consuming and require approximately three to four days to complete a multistep laboratory procedure.

This platform was successfully evaluated for the best sample processing and data analysis procedures and demonstrated a threefold increase in the sequencing output per chip (over 600 Mb compared to 200 Mb), providing superior barcoding options and throughput (up to 32 samples per chip). It should be noted however, that "in-house" alterations of the chip loading protocol, rather than manufacturer recommendations, were mainly responsible for greater sequencing performance in this study. Coupled with frequent updates of the Ion Torrent software, hardware and laboratory protocols, these issues emphasize the current vulnerability of this platform.

The sequencing concordance study results demonstrated a relatively high accuracy of genotyping, which was significantly increased when a higher sequencing depth and stringency of analysis were applied. This pilot study provided novel and promising results for potential forensic use of this platform. However, considering the relatively

low sample size used for this evaluation (due to high cost), an additional study with a larger sample size might be required to further validate these results.

It can be concluded that Ion Torrent with its unprecedented multiplex capacity of genotyping thousands of amplicons and relatively low template input holds promise for forensic application, although it must be extensively validated prior to routine use.

6.7. Genomic association analysis of externally visible traits (EVT) and ancestry

6.7.1. Ancestry association analysis

The majority of candidate craniofacial markers used in this project were selected according to their high Fst values, reflecting high population differentiation. The rationale behind this selection was that different population groups demonstrate distinct differences in facial features. In addition, many pigmentation markers naturally overlap with ancestry informative markers (AIMs) and can potentially predict both characteristics (as demonstrated in this study). This singularity is often used by people (including eyewitness) to estimate individual's ancestry, although being quite an unreliable estimator. Predicting the ancestry of an unknown DNA sample is valuable for both crime investigation, including unidentified human remains, and for anthropology research, including ancient DNA analysis.

This project investigated genetic association of four main ancestries: European, East Asian, African and South Asian (Indian). For European ancestry, the most significantly associated markers were rs16891982 in SLC45A2 gene (p-value 1.01E-66) and rs12913832 in HERC2 gene (p-value 2.34E-27). For Asian ancestry, the most significantly associated markers were rs4823810 in CELSR1 gene (p-value 5.98E-79) and rs12096413 in EIF2C3 gene (p-value 8.36E-70). For African ancestry, the most significantly associated markers were rs2709927 in SEMA3E gene (p-value 3.53E-118) and rs56293475 in RTTN gene (p-value 1.14E-116). For Indian ancestry the most significantly associated markers were rs16891982 in SLC45A2 gene (p-value 3.17E-17) and rs1010872 in SLC45A2 (p-value 4.99E-16). Notably, the European and Indian populations produced associations with similar markers, compared to African and Asian

population groups. This observation was supported by comparison of craniofacial measurements performed in this project as well as by published studies on the linguistic, ethno-geographic and genetic history of the Indo-European population [441, 442].

As expected, some of the significantly associated genes that were identified have a predominant role in the pigmentation process, including HERC2 and SLC45A2, while others such as CELSR1, SEMA3E and RTTN were linked to the craniofacial embryonic development.

In summary, the genetic associations demonstrated in this project were confirmed by previously published results (specifically for the European ancestry), while others provided new associations and potential predictors of ancestry in other population groups.

6.7.2. Pigmentation traits association analysis

Despite the major contributors for differences in eye, skin and hair colour having been identified over the last decade, the function of many genes involved in this process is still not fully understood. Given that known genes only partially explain the normal pigmentation phenotype, identification of other genes involved in this process could provide complimentary information and better prediction accuracy of these traits.

This project has contributed to this goal by both confirming the association with already known genes such as HERC2, OCA2, SLC45A2, SLC24A5 and IRF4 as well as with fifteen novel pigmentation genes, including HDAC4, POU2F3, PCDH15, RPL6, ARHGEF7, DDX3Y, SEMA3E, PTK6, IL20RA and LAMB4 which were not associated previously with particular pigmentation traits, although linked to pigmentation-related cellular process or disease (e.g. melanogenesis or melanoma). In addition, numerous polymorphisms in 38 genes were found to be in strong association with either eye, skin or hair colour, providing new evidence of their potential role in pigmentation. Notably, SNP rs10505261 in the TRPS1 gene was strongly associated with hair curliness (p-value 1.47E-23). Interestingly, this gene was previously associated with the Hypertrichosis syndrome, which involves abnormal amounts of hair growth over the body [450]. However, considering the association of the same gene with dark skin in this study (p-value 7.15E-67), this finding may indicate an insufficient

African population sample size was used. On the other hand, given the very strong association results, these findings may indeed be true positive.

These results not only expand the current knowledge of pigmentation genetics, but also form a basis for incorporating these markers in a pigmentation assay for more robust prediction.

6.7.3. Craniofacial traits association analysis

Genetic association analysis of the craniofacial features included 92 measurements and ratios, collected and calculated using 3D images as well as by direct anthropometric examination using a calliper. All the craniofacial measurements were used to perform a principal component analysis (PCA) and identify 20 principal components that represent all the measurements. Considering the number of measurements used in this project for genetic association analysis, it represents the most comprehensive study performed to date. In addition, the measurements in this project were collected by a single examiner, potentially reducing variability between observations and producing a well-grounded data. Opposite to other published studies, which used a combined dataset of measurements collected from 2D and 3D images of individuals with a wide range of ages, this study focused only on measurements obtained from 3D images of volunteers from a compact age group.

The association analysis was performed using stringent statistical model, incorporating the EIGENSTRAT function for population stratification correction conjointly with covariates such as sex and BMI, potentially reducing the possibility of false positive results. Association analysis revealed 58 genetic markers in 33 different genes and intergenic regions that were significantly associated with thirteen linear distances, four angular distances and one ratio. Interestingly, the nasal area demonstrated the strongest genetic association within the direct measurements, most likely due to the higher reproducibility of the relevant measurements (as discussed in Section 4.4). The most significantly associated markers were found in the following genes: COL11A1, GLI2, LMNA, EPN2 and BMP7 (as shown in Tables 50 and 51). The majority of these genes were previously assigned a role in the craniofacial embryonic development or associated with syndromes displaying various craniofacial abnormalities in humans or model organisms.

An association analysis of a non-anthropometrical phenotypic trait - single vs double eye lid, showed significant association with nine markers in six genes and intergenic regions (Table 52). Several markers in two genes, HYDIN and CELSR1, demonstrated the strongest association with this trait. A single eye lid is especially common in the East Asian population and these SNPs therefore, can also provide ancestry estimation.

The most significant associations were found between principal components and genetic markers, such as rs7531988 in SLC35F3 gene (p-value 8.24E-50), rs4823810 in CELSR1 gene (p-value 5.65E-46) and rs12096413 in EIF2C3 gene (p-value 1.10E-45) (summarized in Table 53). Such a strong association can be explained by the fact that each component (essentially the eigenvector) incorporates numerous measurements, which collectively produce stronger association. The association of eigenvectors with specific genes may illustrate underlying morphological processes during the facial tissues development (for example cells proliferation, polarity and migration), which in essence are regulated by these genes.

All the associations revealed in this study are novel and have not been previously reported. A recently published GWAS found association between the bizygomatic distance (zy-zy) and SNP rs987525 near the CCDC26 gene [160]. While the current study has not investigated this linkage (as the specific marker was not included in the assay), a strong associations between zy-zy distance and a number of markers in the HYDIN gene (lowest p-value 6.76E-06) were identified. Despite that, no functional linkage between these two genes has been found yet. These results provide an additional piece in the complex puzzle of craniofacial genetics.

Overall, the statistical analysis revealed 161 unique markers in 63 genes and intergenic regions associated with either individual craniofacial measurements or principal components, with the majority of markers located in either introns, 5` and 3` of genes or in the intergenic regions. The functional annotation of these markers revealed that 126 of 161 markers may have potentially functional regulatory role, such as affecting transcription binding sites. A search for protein network annotation with all significant genes demonstrated a complex level of interactions between almost all genes (except one).

The identification of novel genetic associations found in this project, provides valuable information for better understanding of the craniofacial embryogenetics and establishes the first attempt to predict facial appearance from a DNA sample.

6.8. Assessment of the prediction power of a SNP set for EVTs and ancestry

Following the genetic association analysis, a pilot study for constructing statistical prediction models was performed. Statistical models, generated for prediction of several phenotypes, including craniofacial traits, demonstrated reasonably high prediction accuracy, specificity and sensitivity. However, an actual prediction of these phenotypes on a "blind" sample set of 25 individuals resulted in lower prediction accuracy. It must be noted, that these results are preliminary and currently undergoing further analysis and refinement.

6.9. Future directions

Genetic regulation of the craniofacial embryonic development and as a result an extraordinary variety in human facial appearance is extremely complex process. This project has briefly investigated a set of candidate markers, which may be responsible for these variations. Numerous genetic markers were found associated with several craniofacial measurements and shapes (represented by principal components), revealing novel and important details of the craniofacial embryogenetics. However, due to time, cost and labour limitations, this study has not achieved its full potential. There is a clear need to continue and extend this study in the near future by implementing several potential directions, such as:

- Developing a robust method to acquire maximum information from the 3D images by collecting a polygonal mesh details, rather than individual landmarks.
- Performing a follow up study with increased sample size of 3D images and DNA samples, especially of specific population groups, such as Africans and Aboriginals. This study is essential to replicate the association results of the current project.
- Performing a full genome sequencing of the same samples to identify additional markers, involved in the craniofacial morphology.
- Performing microRNA and methylation pattern profiling to investigate additional factors that may be involved in this process.

- Conducting an Ion Torrent sensitivity and mixture study to investigate this platform's performance with low DNA quality and quantity as well as its mixture deconvulation potential.

6.10. Final Conclusions

This research project has investigated genetic polymorphisms, potentially influencing normal craniofacial morphology. An application of relatively novel Massively Parallel Sequencing platform for genotyping approximately 6,500 markers enabled to identify numerous markers, significantly associated with specific visible traits. A large database of almost 600 DNA samples and 3D images was collected. In addition to craniofacial characteristics, this study has also investigated an association between known SNPs and pigmentation traits and ancestry. The most significantly associated pigmentation and ancestry markers found in this project were confirmed by previously published results. In addition, a number of novel genes were linked to specific pigmentation traits and population groups. The association of SNPs with European and Asian ancestry was in general consensus with published data, while the identification of most ancestry-informative markers in African and Indian populations were novel.

The association analysis results enabled construction of statistical models for prediction of externally visible traits (EVT) and ancestry. Although preliminary, these results are highly encouraging and demonstrate a first attempt to predict facial appearance and incorporate this information into the all-inclusive forensic Identikit.

References

1. Kolar, J.C. and E.M. Salter, *Craniofaciall Anthropometry: Practical measurement of the head and face for clinical, surgical, and research use*. **1997**: Charles C. Thomas Publisher.

2. Jobling, M.A. and P. Gill, Encoded evidence: DNA in forensic analysis. *Nat Rev Genet*, **2004**. 5(10): p. 739-51.

3. Budowle, B. and A. van Daal, Extracting evidence from forensic DNA analyses: future molecular biology directions. *Biotechniques*, **2009**. 46(5): p. 339-40, 342-50.

4. Wells, G.L. and E.A. Olson, Eyewitness testimony. *Annu Rev Psychol*, **2003**. 54(1): p. 277-95.

5. Kayser, M. and P.M. Schneider, DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations. *Forensic Sci Int Genet*, **2009**. 3(3): p. 154-61.

6. Lindsay, R.C.L., G.L. Wells, and F.J. Oconnor, Mock-Juror Belief of Accurate and Inaccurate Eyewitnesses - a Replication and Extension. *Law and Human Behavior*, **1989**. 13(3): p. 333-339.

7. Wells, G.L., R.S. Malpass, R.C. Lindsay, R.P. Fisher, J.W. Turtle, and S.M. Fulero, From the lab to the police station. A successful application of eyewitness research. *Am Psychol*, **2000**. 55(6): p. 581-98.

8. Mackay, T.F., E.A. Stone, and J.F. Ayroles, The genetics of quantitative traits: challenges and prospects. *Nat Rev Genet*, **2009**. 10(8): p. 565-77.

9. Bouakaze, C., C. Keyser, E. Crubezy, D. Montagnon, and B. Ludes, Pigment phenotype and biogeographical ancestry from ancient skeletal remains: inferences from multiplexed autosomal SNP analysis. *Int J Legal Med*, **2009**. 123(4): p. 315-25.

10. Valenzuela, R.K., M.S. Henderson, M.H. Walsh, N.A. Garrison, J.T. Kelch, O. Cohen-Barak, D.T. Erickson, F. John Meaney, *et al.*, Predicting phenotype from genotype: normal pigmentation. *J Forensic Sci*, **2010**. 55(2): p. 315-22.

11. Branicki, W., F. Liu, K. van Duijn, J. Draus-Barini, E. Pospiech, S. Walsh, T. Kupiec, A. Wojas-Pelc, *et al.*, Model-based prediction of human hair color using DNA variants. *Hum Genet*, **2011**. 129(4): p. 443-54.

12. Bulbul, O., G. Filoglu, H. Altuncul, A.F. Aradas, Y. Ruiz, M. Fondevila, C. Phillips, Á. Carracedo, *et al.*, A SNP multiplex for the simultaneous prediction of biogeographic ancestry and pigmentation type. *Forensic Sci Inter: Genet Suppl*, **2011**. 3(1): p. e500-e501.

13. Ruiz, Y., C. Phillips, A. Gomez-Tato, J. Alvarez-Dios, M. Casares de Cal, R. Cruz, O. Maronas, J. Sochtig, *et al.*, Further development of forensic eye color predictive tests. *Forensic Sci Int Genet*, **2013**. 7(1): p. 28-40.

14. Liu, F., B. Wen, and M. Kayser, Colorful DNA polymorphisms in humans. *Semin Cell Dev Biol*, **2013**. 24(6-7): p. 562-75.

15. Meissner, C.A. and J.C. Brigham, Thirty years of investigating the own-race bias in memory for faces - A meta-analytic review. *Psychology Public Policy and Law*, **2001**. 7(1): p. 3-35.

16. Frudakis, T., *Molecular photofitting: predicting ancestry and phenotype using DNA*. **2010**: Academic Press. 695 p.

17. Farkas, L.G., *Anthropometry of the head and face*. 2nd ed. **1994**, New York: Raven Press. xix, 405 p.

18. Farkas, L.G. and I.R. Monro, *Antropometric facial proportions in medicine*. **1987**: Thomas. 344 p.

19. Garson, J.G., The Frankfort Craniometric Agreement, with Critical Remarks Thereon. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **1885**. 14(Royal Anthropological Institute of Great Britain and Ireland): p. 64-83.

20. Farkas, L.G., T.A. Hreczko, J.C. Kolar, and I.R. Munro, Vertical and horizontal proportions of the face in young adult North American Caucasians: revision of neoclassical canons. *Plast Reconstr Surg*, **1985**. 75(3): p. 328-38.

21. Le, T.T., L.G. Farkas, R.C. Ngim, L.S. Levin, and C.R. Forrest, Proportionality in Asian and North American Caucasian faces using neoclassical facial canons as criteria. *Aesthetic Plast Surg*, **2002**. 26(1): p. 64-9.

22. Wang, D., G. Qian, M. Zhang, and L.G. Farkas, Differences in horizontal, neoclassical facial canons in Chinese (Han) and North American Caucasian populations. *Aesthetic Plast Surg*, **1997**. 21(4): p. 265-9.

23. Holton, N.E., T.R. Yokley, A.W. Froehle, and T.E. Southard, Ontogenetic scaling of the human nose in a longitudinal sample: implications for genus Homo facial evolution. *Am J Phys Anthropol*, **2014**. 153(1): p. 52-60.

24. Haselhuhn, M.P., E.M. Wong, and M.E. Ormiston, Self-fulfilling prophecies as a link between men's facial width-to-height ratio and behavior. *PLoS One*, **2013**. 8(8): p. e72259.

25. Kleisner, K., T. Kočnar, A. Rubešová, and J. Flegr, Eye color predicts but does not directly influence perceived dominance in men. *Personality and Individual Differences*, **2010**. 49(1): p. 59-64.

26. Cakirer, B., M.G. Hans, G. Graham, J. Aylor, P.V. Tishler, and S. Redline, The relationship between craniofacial morphology and obstructive sleep apnea in whites and in African-Americans. *Am J Respir Crit Care Med*, **2001**. 163(4): p. 947-50.

27. Lowe, A.A., J.D. Santamaria, J.A. Fleetham, and C. Price, Facial morphology and obstructive sleep apnea. *Am J Orthod Dentofacial Orthop*, **1986**. 90(6): p. 484-91.

28. Farkas, L.G., M.J. Katic, C.R. Forrest, K.W. Alt, B. I, G. Baltadjiev, E. Cunha, M. Cvicelova, *et al.*, International anthropometric study of facial morphology in various ethnic groups/races. *J Cran Surg*, **2005**. 16(4): p. 615-646.

29. İşcan, M.Y. and R.P. Helmer, *Forensic analysis of the skull: craniofacial analysis, reconstruction, and identification*. **1993**: Wiley-Liss Chichester, NY.

30. Baik, H.S., J.M. Jeon, and H.J. Lee, Facial soft-tissue analysis of Korean adults with normal occlusion using a 3-dimensional laser scanner. *Am J Orthod Dentofacial Orthop*, **2007**. 131(6): p. 759-66.

31. Relethford, J.H., Craniometric variation among modern human populations. *Am J Phys Anthropol*, **1994**. 95(1): p. 53-62.

32. Hennessy, R.J. and C.B. Stringer, Geometric morphometric study of the regional variation of modern human craniofacial form. *Am J Phys Anthropol*, **2002**. 117(1): p. 37-48.

33. Kim, H.-J., S.-W. Im, G. Jargal, S. Lee, J.-H. Yi, J.-Y. Park, J. Sung, S.-I. Cho, *et al.*, Heritabilities of Facial Measurements and Their Latent Factors in Korean Families. *Genomics & informatics*, **2013**. 11(2): p. 83-92.

34. Yadav, A.O., C.S. Walia, R.M. Borle, K.H. Chaoji, R. Rajan, and A.N. Datarkar, Cephalometric norms for Central Indian population using Burstone and Legan analysis. *Indian J Dent Res*, **2011**. 22(1): p. 28-33.

35. Hrdlička, A., Anthropometry. *Am J Phys Anthropol*, **1920**. 3(1): p. 147-173.

36. Douglas, T.S., Image processing for craniofacial landmark identification and measurement: a review of photogrammetry and cephalometry. *Comput Med Imaging Graph*, **2004**. 28(7): p. 401-9.

37. Heike, C.L., K. Upson, E. Stuhaug, and S.M. Weinberg, 3D digital stereophotogrammetry: a practical guide to facial image acquisition. *Head Face Med*, **2010**. 6(1): p. 18.

38. Kau, C.H., A. Zhurov, R. Bibb, L. Hunter, and S. Richmond, The investigation of the changing facial appearance of identical twins employing a three-dimensional laser imaging system. *Orthod Craniofac Res*, **2005**. 8(2): p. 85-90.

39. Kau, C.H., A. Zhurov, S. Richmond, A. Cronin, C. Savio, and C. Mallorie, Facial templates: a new perspective in three dimensions. *Orthod Craniofac Res*, **2006**. 9(1): p. 10-7.

40. Schmidt, E.J., T.E. Parsons, H.A. Jamniczky, J. Gitelman, C. Trpkov, J.C. Boughner, C.C. Logan, C.W. Sensen, *et al.*, Micro-computed tomography-based phenotypic approaches in embryology: procedural artifacts on assessments of embryonic craniofacial growth and development. *BMC Dev Biol*, **2010**. 10: p. 18.

41. Toma, A.M., A. Zhurov, R. Playle, E. Ong, and S. Richmond, Reproducibility of facial soft tissue landmarks on 3D laser-scanned facial images. *Orthod Craniofac Res*, **2009**. 12(1): p. 33-42.

42. Toma, A.M., A. Zhurov, R. Playle, and S. Richmond, A three-dimensional look for facial differences between males and females in a British-Caucasian sample aged 151/2 years old. *Orthod Craniofac Res*, **2008**. 11(3): p. 180-5.

43. Weinberg, S.M., S.D. Naidoo, K.M. Bardi, C.A. Brandon, K. Neiswanger, J.M. Resick, R.A. Martin, and M.L. Marazita, Face shape of unaffected parents with cleft affected offspring: combining three-dimensional surface imaging and geometric morphometrics. *Orthod Craniofac Res*, **2009**. 12(4): p. 271-81.

44. Pan Zheng, Bahari Belaton, Rozniza Zaharudin, and A. Irani, Computerized 3D Craniofacial Landmark Identification and Analysis. *eJCSIT Electronic Journal of Computer Science & Information Technology*, **2009**. 1(1).

45. Hammond, P., T.J. Hutton, J.E. Allanson, L.E. Campbell, R.C. Hennekam, S. Holden, M.A. Patton, A. Shaw, *et al.*, 3D analysis of facial morphology. *Am J Med Genet A*, **2004**. 126A(4): p. 339-48.

46. Sholts, S.B., S.K. Warmlander, L.M. Flores, K.W. Miller, and P.L. Walker, Variation in the measurement of cranial volume and surface area using 3D laser scanning technology. *J Forensic Sci*, **2010**. 55(4): p. 871-6.

47. Gupta, S., M.K. Markey, and A.C. Bovik, Anthropometric 3D Face Recognition. *Int J Comput Vision*, **2010**. 90(3): p. 331-349.

48. Gupta, S., K. Castleman, M. Markey, and A. Bovik. Texas 3D face recognition database. in *Image Analysis & Interpretation (SSIAI), 2010 IEEE Southwest Symposium on*. **2010**. *IEEE*.

49. Koo, H.S. and K.M. Lam, Recovering the 3D shape and poses of face images based on the similarity transform. *Pattern Recogn Lett*, **2008**. 29(6): p. 712-723.

50. Fedosyutkin, B.A. and J.V. Nainys, The relationship of skull morphology to facial features. *Forensic analysis of the skull. New York, NY: Wiley-Liss*, **1993**: p. 199-213.

51. Quatrehomme, G. and M. Isçan, Computerized facial reconstruction. *Encyclopedia of forensic sciences. Academic, San Diego*, **2000**: p. 773-779.

52. Aulsebrook, W.A., M.Y. Iscan, J.H. Slabbert, and P. Becker, Superimposition and reconstruction in forensic facial identification: a survey. *Forensic Sci Int*, **1995**. 75(2-3): p. 101-20.

53. Tyrrell, A.J., M.P. Evison, A.T. Chamberlain, and M.A. Green, Forensic three-dimensional facial reconstruction: historical review and contemporary developments. *J Forensic Sci*, **1997**. 42(4): p. 653-61.

54. Wilkinson, C., *Forensic facial reconstruction*. **2004**: Cambridge University Press. 292pp.

55. Vanezis, P., M. Vanezis, G. McCombe, and T. Niblett, Facial reconstruction using 3-D computer graphics. *Forensic Sci Int*, **2000**. 108(2): p. 81-95.

56. Wilkinson, C., Computerized forensic facial reconstruction. *Foren Sci Med Path*, **2005**. 1(3): p. 173-177.

57. Bruder, C.E., A. Piotrowski, A.A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Stahl, U. Menzel, J. Sandgren, *et al.*, Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet*, **2008**. 82(3): p. 763-71.

58. Fraga, M.F., E. Ballestar, M.F. Paz, S. Ropero, F. Setien, M.L. Ballestar, D. Heine-Suner, J.C. Cigudosa, *et al.*, Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci U S A*, **2005**. 102(30): p. 10604-9.

59. Halder, A., M. Jain, I. Chaudhary, and B. Varma, Chromosome 22q11.2 microdeletion in monozygotic twins with discordant phenotype and deletion size. *Mol Cytogenet*, **2012**. 5(1): p. 13.

60. Kaminsky, Z.A., T. Tang, S.C. Wang, C. Ptak, G.H. Oh, A.H. Wong, L.A. Feldcamp, C. Virtanen, *et al.*, DNA methylation profiles in monozygotic and dizygotic twins. *Nat Genet*, **2009**. 41(2): p. 240-5.

61. Nilsson, L. A child is born. [Internet] **2013** [cited 2014 March]; Available from: http://www.lennartnilsson.com/child_is_born.html.

62. Sperber, G.H., S.M. Sperber, and G.D. Guttmann, *Craniofacial embryogenetics and development*. 2nd ed. **2010**, Shelton, CT: People's Medical Pub. House USA. 250p.

63. Jill A. Helms, D.C., Minal D. Tapadia, New insights into craniofacial morphogenesis. *Development*, **2005**. 132: p. 851-861.

64. Tapadia, M.D., D.R. Cordero, and J.A. Helms, It's all in your head: new insights into craniofacial development and deformation. *J. Anat.*, **2005**. 207: p. 461–477.

65. Schneider, R.A., D. Hu, and J.A. Helms, From head to toe: conservation of molecular signals regulating limb and craniofacial morphogenesis. *Cell Tissue Res*, **1999**. 296(1): p. 103-9.

66. Roessler, E., E. Belloni, K. Gaudenz, P. Jay, P. Berta, S.W. Scherer, L.C. Tsui, and M. Muenke, Mutations in the human Sonic Hedgehog gene cause holoprosencephaly. *Nat Genet*, **1996**. 14(3): p. 357-60.

67. Mavrogiannis, L.A., I. Antonopoulou, A. Baxova, S. Kutilek, C.A. Kim, S.M. Sugayama, A. Salamanca, S.A. Wall, *et al.*, Haploinsufficiency of the human

homeobox gene ALX4 causes skull ossification defects. *Nat Genet*, **2001**. 27(1): p. 17-8.

68. Qu, S., S.C. Tucker, J.S. Ehrlich, J.M. Levorse, L.A. Flaherty, R. Wisdom, and T.F. Vogt, Mutations in mouse Aristaless-like4 cause Strong's luxoid polydactyly. *Development*, **1998**. 125(14): p. 2711-21.

69. Begum, F., D. Ghosh, G.C. Tseng, and E. Feingold, Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res*, **2012**. 40(9): p. 3777-84.

70. Manolio, T.A., F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, *et al.*, Finding the missing heritability of complex diseases. *Nature*, **2009**. 461(7265): p. 747-53.

71. Hindorff, L.A., P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, and T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, **2009**. 106(23): p. 9362-7.

72. Rees, J.L., Genetics of hair and skin color. *Annu Rev Genet*, **2003**. 37(1): p. 67-90.

73. Yamaguchi, T., K. Maki, and Y. Shibasaki, Growth hormone receptor gene variant and mandibular height in the normal Japanese population. *Am J Orthod Dentofacial Orthop*, **2001**. 119(6): p. 650-3.

74. Coussens, A.K. and A. van Daal, Linkage disequilibrium analysis identifies an FGFR1 haplotype-tag SNP associated with normal variation in craniofacial shape. *Genomics*, **2005**. 85(5): p. 563-73.

75. Ermakov, S., M.G. Rosenbaum, I. Malkin, and G. Livshits, Family-based study of association between ENPP1 genetic variants and craniofacial morphology. *Ann Hum Biol*, **2010**. 37(6): p. 754-66.

76. Paternoster, L., A.I. Zhurov, A.M. Toma, J.P. Kemp, B. St Pourcain, N.J. Timpson, G. McMahon, W. McArdle, *et al.*, Genome-wide association study of

three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *Am J Hum Genet*, **2012**. 90(3): p. 478-85.

77. Liu, F., F. van der Lijn, C. Schurmann, G. Zhu, M.M. Chakravarty, P.G. Hysi, A. Wollstein, O. Lao, *et al.*, A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet*, **2012**. 8(9): p. e1002932.

78. Berndt, S.I., S. Gustafsson, R. Magi, A. Ganna, E. Wheeler, M.F. Feitosa, A.E. Justice, K.L. Monda, *et al.*, Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet*, **2013**. 45(5): p. 501-12.

79. Wilkie, A.O. and G.M. Morriss-Kay, Genetics of craniofacial development and malformation. *Nat Rev Genet*, **2001**. 2(6): p. 458-68.

80. Kimonis, V., J.A. Gold, T.L. Hoffman, J. Panchal, and S.A. Boyadjiev, Genetics of craniosynostosis. *Semin Pediatr Neurol*, **2007**. 14(3): p. 150-61.

81. Johnson, D. and A.O. Wilkie, Craniosynostosis. *Eur J Hum Genet*, **2011**. 19(4): p. 369-76.

82. Derderian, C. and J. Seaward, Syndromic craniosynostosis. *Semin Plast Surg*, **2012**. 26(2): p. 64-75.

83. Garza, R.M. and R.K. Khosla, Nonsyndromic craniosynostosis. *Semin Plast Surg*, **2012**. 26(2): p. 53-63.

84. Lindsay, E.A., A. Botta, V. Jurecic, S. Carattini-Rivera, Y.C. Cheah, H.M. Rosenblatt, A. Bradley, and A. Baldini, Congenital heart disease in mice deficient for the DiGeorge syndrome region. *Nature*, **1999**. 401(6751): p. 379-83.

85. Scambler, P.J., The 22q11 deletion syndromes. *Hum Mol Genet*, **2000**. 9(16): p. 2421-6.

86. Jerome, L.A. and V.E. Papaioannou, DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. *Nat Genet*, **2001**. 27(3): p. 286-91.

87. Merscher, S., B. Funke, J.A. Epstein, J. Heyer, A. Puech, M.M. Lu, R.J. Xavier, M.B. Demay, *et al.*, TBX1 is responsible for cardiovascular defects in Velo-Cardio-Facial/DiGeorge syndrome. *Cell*, **2001**. 104(4): p. 619-629.

88. Aggarwal, V.S., C. Carpenter, L. Freyer, J. Liao, M. Petti, and B.E. Morrow, Mesodermal Tbx1 is required for patterning the proximal mandible in mice. *Dev Biol*, **2010**. 344(2): p. 669-81.

89. Guo, C., Y. Sun, B. Zhou, R.M. Adam, X. Li, W.T. Pu, B.E. Morrow, A. Moon, *et al.*, A Tbx1-Six1/Eya1-Fgf8 genetic pathway controls mammalian cardiovascular and craniofacial morphogenesis. *J Clin Invest*, **2011**. 121(4): p. 1585-95.

90. el Ghouzzi, V., M. Le Merrer, F. Perrin-Schmitt, E. Lajeunie, P. Benit, D. Renier, P. Bourgeois, A.L. Bolcato-Bellemin, *et al.*, Mutations of the TWIST gene in the Saethre-Chotzen syndrome. *Nat Genet*, **1997**. 15(1): p. 42-6.

91. Howard, T.D., W.A. Paznekas, E.D. Green, L.C. Chiang, N. Ma, R.I. Ortiz de Luna, C. Garcia Delgado, M. Gonzalez-Ramos, *et al.*, Mutations in TWIST, a basic helix-loop-helix transcription factor, in Saethre-Chotzen syndrome. *Nat Genet*, **1997**. 15(1): p. 36-41.

92. Bourgeois, P., A.L. Bolcato-Bellemin, J.M. Danse, A. Bloch-Zupan, K. Yoshiba, C. Stoetzel, and F. Perrin-Schmitt, The variable expressivity and incomplete penetrance of the twist-null heterozygous mouse phenotype resemble those of human Saethre-Chotzen syndrome. *Hum Mol Genet*, **1998**. 7(6): p. 945-957.

93. El Ghouzzi, V., L. Legeai-Mallet, C. Benoist-Lasselin, E. Lajeunie, D. Renier, A. Munnich, and J. Bonaventure, Mutations in the basic domain and the loop-helix II junction of TWIST abolish DNA binding in Saethre-Chotzen syndrome. *FEBS Lett*, **2001**. 492(1-2): p. 112-8.

94. Wuyts, W., E. Cleiren, T. Homfray, A. Rasore-Quartino, F. Vanhoenacker, and W. Van Hul, The ALX4 homeobox gene is mutated in patients with ossification defects of the skull (foramina parietalia permagna, OMIM 168500). *J Med Genet*, **2000**. 37(12): p. 916-920.

95. Muenke, M., U. Schell, A. Hehr, N.H. Robin, H.W. Losken, A. Schinzel, L.J. Pulleyn, P. Rutland, *et al.*, A common mutation in the fibroblast growth factor receptor 1 gene in Pfeiffer syndrome. *Nat Genet*, **1994**. 8(3): p. 269-74.

96. Glaser, R.L., W. Jiang, S.A. Boyadjiev, A.K. Tran, A.A. Zachary, L. Van Maldergem, D. Johnson, S. Walsh, *et al.*, Paternal origin of FGFR2 mutations in sporadic cases of Crouzon syndrome and Pfeiffer syndrome. *Am J Hum Genet*, **2000**. 66(3): p. 768-77.

97. Nishimura, G., N. Haga, H. Kitoh, Y. Tanaka, T. Sonoda, M. Kitamura, S. Shirahama, T. Itoh, *et al.*, The phenotypic spectrum of COL2A1 mutations. *Human mutation*, **2005**. 26(1): p. 36-43.

98. Baas, D., M. Malbouyres, Z. Haftek-Terreau, D. Le Guellec, and F. Ruggiero, Craniofacial cartilage morphogenesis requires zebrafish col11a1 activity. *Matrix Biol*, **2009**. 28(8): p. 490-502.

99. Annunen, S., J. Korkko, M. Czarny, M.L. Warman, H.G. Brunner, H. Kaariainen, J.B. Mulliken, L. Tranebjaerg, *et al.*, Splicing mutations of 54-bp exons in the COL11A1 gene cause Marshall syndrome, but other mutations cause overlapping Marshall/Stickler phenotypes. *Am J Hum Genet*, **1999**. 65(4): p. 974-83.

100. Richards, A.J., J.R.W. Yates, R. Williams, S.J. Payne, F.M. Pope, J.D. Scott, and M.P. Snead, A family with Stickler syndrome type 2 has a mutation in the COL11A1 gene resulting in the substitution of glycine 97 by valine in alpha 1(XI) collagen. *Human Molecular Genetics*, **1996**. 5(9): p. 1339-1343.

101. Farkas, L.G., M.J. Katic, and C.R. Forrest, Surface anatomy of the face in Down's syndrome: anthropometric proportion indices in the craniofacial regions. *J Craniofac Surg*, **2001**. 12(6): p. 519-24; discussion 525-6.

102. Ferrario, V.F., C. Dellavia, G. Serrao, and C. Sforza, Soft tissue facial angles in Down's syndrome subjects: a three-dimensional non-invasive study. *Eur J Orthod*, **2005**. 27(4): p. 355-62.

103. Sforza, C., C. Dellavia, C. Dolci, E. Donetti, and V.F. Ferrario, A quantitative three-dimensional assessment of abnormal variations in the facial soft tissues of individuals with Down syndrome. *Cleft Palate Craniofac J*, **2005**. 42(4): p. 410-6.

104. Sforza, C., G. Grandi, L. Pisoni, C. Di Blasio, M. Gandolfini, and V.F. Ferrario, Soft tissue facial morphometry in subjects with Moebius syndrome. *Eur J Oral Sci*, **2009**. 117(6): p. 695-703.

105. Starbuck, J., R.H. Reeves, and J. Richtsmeier, Morphological integration of soft-tissue facial morphology in Down Syndrome and siblings. *Am J Phys Anthropol*, **2011**. 146(4): p. 560-8.

106. Sforza, C., C. Dellavia, C. Allievi, D.G. Tommasi, and V.F. Ferrario, *Anthropometric Indices of Facial Features in Down's Syndrome Subjects*, in Handbook of Anthropometry. **2012**, Springer. p. 1603-1618.

107. Sforza, C., F. Elamin, C. Dellavia, R. Rosati, G. Lodetti, A. Mapelli, and V.F. Ferrario, Morphometry of the orbital region soft tissues in Down syndrome. *J Craniofac Surg*, **2012**. 23(1): p. 198-202.

108. Sybert, V.P. and E. McCauley, Turner's syndrome. *N Engl J Med*, **2004**. 351(12): p. 1227-38.

109. Scutt, L.E., E.W. Chow, R. Weksberg, W.G. Honer, and A.S. Bassett, Patterns of dysmorphic features in schizophrenia. *Am J Med Genet*, **2001**. 105(8): p. 713-23.

110. Hennessy, R.J., P.A. Baldwin, D.J. Browne, A. Kinsella, and J.L. Waddington, Three-dimensional laser surface imaging and geometric morphometrics resolve frontonasal dysmorphology in schizophrenia. *Biol Psychiatry*, **2007**. 61(10): p. 1187-94.

111. Hennessy, R.J., S. McLearie, A. Kinsella, and J.L. Waddington, Facial surface analysis by 3D laser scanning and geometric morphometrics in relation to sexual dimorphism in cerebral--craniofacial morphogenesis and cognitive function. *J Anat*, **2005**. 207(3): p. 283-95.

112. Xu, B., I. Ionita-Laza, J.L. Roos, B. Boone, S. Woodrick, Y. Sun, S. Levy, J.A. Gogos, *et al.*, De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet*, **2012**. 44(12): p. 1365-9.

113. Murphy, K.C., L.A. Jones, and M.J. Owen, High rates of schizophrenia in adults with velo-cardio-facial syndrome. *Arch Gen Psychiatry*, **1999**. 56(10): p. 940-5.

114. Orrico, A., L. Galli, M.L. Cavaliere, L. Garavelli, J.-P. Fryns, E. Crushell, M.M. Rinaldi, A. Medeira, *et al.*, Phenotypic and molecular characterisation of the Aarskog-Scott syndrome: a survey of the clinical variability in light of FGD1 mutation analysis in 46 patients. *Eur J Hum Genet*, **2003**. 12(1): p. 16-23.

115. Kamath, B.M., C. Stolle, L. Bason, R.P. Colliton, D.A. Piccoli, N.B. Spinner, and I.D. Krantz, Craniosynostosis in Alagille syndrome. *Am J Med Genet*, **2002**. 112(2): p. 176-80.

116. Kawara, H., T. Yamamoto, N. Harada, K. Yoshiura, N. Niikawa, A. Nishimura, T. Mizuguchi, and N. Matsumoto, Narrowing candidate region for monosomy 9p syndrome to a 4.7-Mb segment at 9p22.2-p23. *Am J Med Genet A*, **2006**. 140(4): p. 373-7.

117. Andreou, A., A. Lamy, V. Layet, D. Cailliez, F. Gobet, C. Pfister, M. Menard, and T. Frebourg, Early-onset low-grade papillary carcinoma of the bladder associated with Apert syndrome and a germline FGFR2 mutation (Pro253Arg). *Am J Med Genet A*, **2006**. 140(20): p. 2245-7.

118. Weksberg, R., C. Shuman, O. Caluseriu, A.C. Smith, Y.L. Fei, J. Nishikawa, T.L. Stockley, L. Best, *et al.*, Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Hum Mol Genet*, **2002**. 11(11): p. 1317-25.

119. Hatada, I., H. Ohashi, Y. Fukushima, Y. Kaneko, M. Inoue, Y. Komoto, A. Okada, S. Ohishi, *et al.*, An imprinted gene p57KIP2 is mutated in Beckwith–Wiedemann syndrome. *Nat Gen*, **1996**. 14(2): p. 171-173.

120. Catchpoole, D., A.V. Smallwood, J.A. Joyce, A. Murrell, W. Lam, T. Tang, D. Munroe, W. Reik, *et al.*, Mutation analysis of H19 and NAP1L4 (hNAP2)

candidate genes and IGF2 DMR2 in Beckwith-Wiedemann syndrome. *J Med Genet*, **2000**. 37(3): p. 212-5.

121. Kolehmainen, J., R. Wilkinson, A.E. Lehesjoki, K. Chandler, S. Kivitie-Kallio, J. Clayton-Smith, A.L. Traskelin, L. Waris, *et al.*, Delineation of Cohen syndrome following a large-scale genotype-phenotype screen. *Am J Hum Genet*, **2004**. 75(1): p. 122-7.

122. Wu, Q., E. Niebuhr, H. Yang, and L. Hansen, Determination of the 'critical region' for cat-like cry of Cri-du-chat syndrome and analysis of candidate genes by quantitative PCR. *Eur J Hum Genet*, **2005**. 13(4): p. 475-85.

123. Reardon, W., R.M. Winter, P. Rutland, L.J. Pulleyn, B.M. Jones, and S. Malcolm, Mutations in the fibroblast growth factor receptor 2 gene cause Crouzon syndrome. *Nat Genet*, **1994**. 8(1): p. 98-103.

124. Nakamura, A., M. Hattori, and Y. Sakaki, A novel gene isolated from human placenta located in Down syndrome critical region on chromosome 21. *DNA Res*, **1997**. 4(5): p. 321-4.

125. Roper, R.J., L.L. Baxter, N.G. Saran, D.K. Klinedinst, P.A. Beachy, and R.H. Reeves, Defective cerebellar response to mitogenic Hedgehog signaling in Down's syndrome mice. *Proc Natl Acad Sci U S A*, **2006**. 103(5): p. 1452-1456.

126. Hood, R.L., M.A. Lines, S.M. Nikkel, J. Schwartzentruber, C. Beaulieu, M.J. Nowaczyk, J. Allanson, C.A. Kim, *et al.*, Mutations in SRCAP, encoding SNF2-related CREBBP activator protein, cause Floating-Harbor syndrome. *Am J Hum Genet*, **2012**. 90(2): p. 308-13.

127. Machado-Ferreira Mdo, C., M.A. Costa-Lima, R.T. Boy, G.S. Esteves, and M.M. Pimentel, Premature ovarian failure and FRAXA premutation: Positive correlation in a Brazilian survey. *Am J Med Genet A*, **2004**. 126A(3): p. 237-40.

128. Chonchaiya, W., J. Au, A. Schneider, D. Hessl, S.W. Harris, M. Laird, Y. Mu, F. Tassone, *et al.*, Increased prevalence of seizures in boys who were probands with the FMR1 premutation and co-morbid autism spectrum disorder. *Hum Genet*, **2012**. 131(4): p. 581-9.

129. McBrien, J., J.A. Crolla, S. Huang, J. Kelleher, J. Gleeson, and S.A. Lynch, Further case of microdeletion of 8q24 with phenotype overlapping Langer-Giedion without TRPS1 deletion. *Am J Med Genet A*, **2008**. 146A(12): p. 1587-92.

130. Croonen, E.A., I. van der Burgt, L. Kapusta, and J.M. Draaisma, Electrocardiography in Noonan syndrome PTPN11 gene mutation--phenotype characterization. *Am J Med Genet A*, **2008**. 146(3): p. 350-3.

131. Kondoh, T., E. Ishii, Y. Aoki, T. Shimizu, M. Zaitsu, Y. Matsubara, and H. Moriuchi, Noonan syndrome with leukaemoid reaction and overproduction of catecholamines: a case report. *Eur J Pediatr*, **2003**. 162(7-8): p. 548-9.

132. Izumi, K., L.K. Conlin, D. Berrodin, C. Fincher, A. Wilkens, C. Haldeman-Englert, S.C. Saitta, E.H. Zackai, *et al.*, Duplication 12p and Pallister-Killian syndrome: a case report and review of the literature toward defining a Pallister-Killian syndrome minimal critical region. *Am J Med Genet A*, **2012**. 158A(12): p. 3033-45.

133. Jønch, A.E., L.G. Larsen, S. Pouplier, K. Nielsen, K. Brøndum‐Nielsen, and Z. Tümer, Partial duplication of 13q31. 3–q34 and deletion of 13q34 associated with diaphragmatic hernia as a sole malformation in a fetus. *Am J Med Genet A*, **2012**. 158(9): p. 2302-2308.

134. Bellus, G.A., K. Gaudenz, E.H. Zackai, L.A. Clarke, J. Szabo, C.A. Francomano, and M. Muenke, Identical mutations in three different fibroblast growth factor receptor genes in autosomal dominant craniosynostosis syndromes. *Nat Genet*, **1996**. 14(2): p. 174-6.

135. Slager, R.E., T.L. Newton, C.N. Vlangos, B. Finucane, and S.H. Elsea, Mutations in RAI1 associated with Smith-Magenis syndrome. *Nat Genet*, **2003**. 33(4): p. 466-8.

136. Girirajan, S., L.J. Elsas, 2nd, K. Devriendt, and S.H. Elsea, RAI1 variations in Smith-Magenis syndrome patients without 17p11.2 deletions. *J Med Genet*, **2005**. 42(11): p. 820-8.

137. Hunemeier, T., F.M. Salzano, and M.C. Bortolini, TCOF1 T/Ser variant and brachycephaly in dogs. *Anim Genet*, **2009**. 40(3): p. 357-8.

138. Masotti, C., L.M. Armelin-Correa, A. Splendore, C.J. Lin, A. Barbosa, M.C. Sogayar, and M.R. Passos-Bueno, A functional SNP in the promoter region of TCOF1 is associated with reduced gene expression and YY1 DNA-protein interaction. *Gene*, **2005**. 359: p. 44-52.

139. Haworth, K.E., I. Islam, M. Breen, W. Putt, E. Makrinou, M. Binns, D. Hopkinson, and Y. Edwards, Canine TCOF1; cloning, chromosome assignment and genetic analysis in dogs with different head types. *Mamm Genome*, **2001**. 12(8): p. 622-9.

140. Meyer, T. and M.N. Teruel, Fluorescence imaging of signaling networks. *Trends Cell Biol*, **2003**. 13(2): p. 101-6.

141. Clement-Jones, M., S. Schiller, E. Rao, R.J. Blaschke, A. Zuniga, R. Zeller, S.C. Robson, G. Binder, *et al.*, The short stature homeobox gene SHOX is involved in skeletal abnormalities in Turner syndrome. *Hum Mol Genet*, **2000**. 9(5): p. 695-702.

142. Yagi, H., Y. Furutani, H. Hamada, T. Sasaki, S. Asakawa, S. Minoshima, F. Ichida, K. Joo, *et al.*, Role of TBX1 in human del22q11.2 syndrome. *Lancet*, **2003**. 362(9393): p. 1366-73.

143. Hoth, C.F., A. Milunsky, N. Lipsky, R. Sheffer, S.K. Clarren, and C.T. Baldwin, Mutations in the Paired Domain of the Human Pax3 Gene Cause Klein-Waardenburg Syndrome (Ws-Iii) as Well as Waardenburg Syndrome Type-I (Ws-I). *Am J Hum Genet*, **1993**. 52(3): p. 455-462.

144. Mangold, E., K.U. Ludwig, and M.M. Nothen, Breakthroughs in the genetics of orofacial clefting. *Trends Mol Med*, **2011**. 17(12): p. 725-33.

145. Dixon, M.J., M.L. Marazita, T.H. Beaty, and J.C. Murray, Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet*, **2011**. 12(3): p. 167-78.

146. Schutte, B.C. and J.C. Murray, The many faces and factors of orofacial clefts. *Hum Mol Genet*, **1999**. 8(10): p. 1853-9.

147. Lidral, A.C., Association of MSX1 and TGF[beta]3 with nonsyndromic clefting in humans. *Am. J. Hum. Genet.*, **1998**. 63: p. 557-568.

148. Shi, J., X. Jiao, T. Song, B. Zhang, C. Qin, and F. Cao, CRISPLD2 polymorphisms are associated with non-syndromic cleft lip with or without cleft palate in a northern Chinese population. *Eur J Oral Sci*, **2010**. 118(4): p. 430-3.

149. Letra, A., R. Menezes, M.E. Cooper, R.F. Fonseca, S. Tropp, M. Govil, J.M. Granjeiro, S.R. Imoehl, *et al.*, CRISPLD2 variants including a C471T silent mutation may contribute to nonsyndromic cleft lip with or without cleft palate. *Cleft Palate Craniofac J*, **2011**. 48(4): p. 363-70.

150. Rojas-Martinez, A., H. Reutter, O. Chacon-Camacho, R.B. Leon-Cachon, S.G. Munoz-Jimenez, S. Nowak, J. Becker, R. Herberz, *et al.*, Genetic risk factors for nonsyndromic cleft lip with or without cleft palate in a Mesoamerican population: Evidence for IRF6 and variants at 8q24 and 10q25. *Birth Defects Res A Clin Mol Teratol*, **2010**. 88(7): p. 535-7.

151. Marazita, M.L., A.C. Lidral, J.C. Murray, L.L. Field, B.S. Maher, T. Goldstein McHenry, M.E. Cooper, M. Govil, *et al.*, Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. *Hum Hered*, **2009**. 68(3): p. 151-70.

152. Beaty, T.H., J.C. Murray, M.L. Marazita, R.G. Munger, I. Ruczinski, J.B. Hetmanski, K.Y. Liang, T. Wu, *et al.*, A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet*. 42(6): p. 525-529.

153. Mangold, E., K.U. Ludwig, S. Birnbaum, C. Baluardo, M. Ferrian, S. Herms, H. Reutter, N.A. de Assis, *et al.*, Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet*, **2010**. 42(1): p. 24-26.

154. Prescott, N.J., M.M. Lees, R.M. Winter, and S. Malcolm, Identification of susceptibility loci for nonsyndromic cleft lip with or without cleft palate in a two stage genome scan of affected sib-pairs. *Hum Genet*, **2000**. 106(3): p. 345-350.

155. Murray, J., Gene/environment causes of cleft lip and/or palate. *Clin Genet*, **2002**. 61(4): p. 248-256.

156. Christensen, K. and P. Fogh‐Andersen, Cleft lip (±cleft palate) in Danish twins, 1970–1990. *Am J Med Genet*, **1993**. 47(6): p. 910-916.

157. Weinberg, S.M., B.S. Maher, and M.L. Marazita, Parental craniofacial morphology in cleft lip with or without cleft palate as determined by cephalometry: a meta-analysis. *Orthod Craniofac Res*, **2006**. 9(1): p. 18-30.

158. Muenke, M., The pit, the cleft and the web. *Nature genetics*, **2002**. 32(2): p. 219-220.

159. Weinberg, S.M., K. Neiswanger, J.T. Richtsmeier, B.S. Maher, M.P. Mooney, M.I. Siegel, and M.L. Marazita, Three-dimensional morphometric analysis of craniofacial shape in the unaffected relatives of individuals with nonsyndromic orofacial clefts: a possible marker for genetic susceptibility. *Am J Med Genet A*, **2008**. 146A(4): p. 409-20.

160. Boehringer, S., F. van der Lijn, F. Liu, M. Gunther, S. Sinigerova, S. Nowak, K.U. Ludwig, R. Herberz, *et al.*, Genetic determination of human facial morphology: links between cleft-lips and normal variation. *Eur J Hum Genet*, **2011**. 19(11): p. 1192-7.

161. Hochheiser, H., B.J. Aronow, K. Artinger, T.H. Beaty, J.F. Brinkley, Y. Chai, D. Clouthier, M.L. Cunningham, *et al.*, The FaceBase Consortium: a comprehensive program to facilitate craniofacial research. *Dev Biol*, **2011**. 355(2): p. 175-82.

162. Thyagarajan, T., S. Totey, M.J. Danton, and A.B. Kulkarni, Genetically altered mouse models: the good, the bad, and the ugly. *Crit Rev Oral Biol Med*, **2003**. 14(3): p. 154-74.

163. Francis-West, P.H., L. Robson, and D.J.R. Evans, *Craniofacial development : the tissue and molecular interactions that control development of the head*. Advances in anatomy, embryology, and cell biology. **2003**, Berlin ; New York: Springer. vi, 144 p.

164. Mouse Mutant Resource Web Site, The Jackson Laboratory. **2010** [cited 2013 March]; Available from: http://mousemutant.jax.org/.

165. Rossel, M. and M.R. Capecchi, Mice mutant for both Hoxa1 and Hoxb1 show extensive remodeling of the hindbrain and defects in craniofacial development. *Development*, **1999**. 126(22): p. 5027-5040.

166. Brand, M., C.P. Heisenberg, R.M. Warga, F. Pelegri, R.O. Karlstrom, D. Beuchle, A. Picker, Y.J. Jiang, *et al.*, Mutations affecting development of the midline and general body shape during zebrafish embryogenesis. *Development*, **1996**. 123: p. 129-42.

167. Eames, B.F., A. Singer, G.A. Smith, Z.A. Wood, Y.L. Yan, X. He, S.J. Polizzi, J.M. Catchen, *et al.*, UDP xylose synthase 1 is required for morphogenesis and histogenesis of the craniofacial skeleton. *Dev Biol*, **2010**. 341(2): p. 400-15.

168. Miller, C.T., T.F. Schilling, K. Lee, J. Parker, and C.B. Kimmel, sucker encodes a zebrafish Endothelin-1 required for ventral pharyngeal arch development. *Development*, **2000**. 127(17): p. 3815-28.

169. Qi, H.H., M. Sarkissian, G.Q. Hu, Z. Wang, A. Bhattacharjee, D.B. Gordon, M. Gonzales, F. Lan, *et al.*, Histone H4K20/H3K9 demethylase PHF8 regulates zebrafish brain and craniofacial development. *Nature*, **2010**. 466(7305): p. 503-7.

170. van Boxtel, A.L., B. Pieterse, P. Cenijn, J.H. Kamstra, A. Brouwer, W. van Wieringen, J. de Boer, and J. Legler, Dithiocarbamates induce craniofacial abnormalities and downregulate sox9a during zebrafish development. *Toxicol Sci*, **2010**. 117(1): p. 209-17.

171. Yelick, P.C. and T.F. Schilling, Molecular dissection of craniofacial development using zebrafish. *Crit Rev Oral Biol Med*, **2002**. 13(4): p. 308-322.

172. Terai, Y., N. Morikawa, and N. Okada, The evolution of the pro-domain of bone morphogenetic protein 4 (Bmp4) in an explosively speciated lineage of East African cichlid fishes. *Mol Biol Evol*, **2002**. 19(9): p. 1628-32.

173. Albertson, R.C., J.T. Streelman, T.D. Kocher, and P.C. Yelick, Integration and evolution of the cichlid mandible: the molecular basis of alternate feeding strategies. *Proc Natl Acad Sci U S A*, **2005**. 102(45): p. 16287-92.

174. Hamamori, Y., Regulation of histone acetyltranferases p300 and PCAF by the bHLH protein Twist and adenoviral oncoprotein E1A. *Cell*, **1999**. 96: p. 405-413.

175. Johnson, D., A comprehensive screen for TWISTmutations in patients with craniosynostosis identifies a new microdeletion syndrome of chromosome band 7p21.1. *Am. J. Hum. Genet.*, **1998**. 63: p. 1282-1293.

176. Johnson, D., S. Iseki, A.O. Wilkie, and G.M. Morriss-Kay, Expression patterns of Twist and Fgfr1, -2 and -3 in the developing mouse coronal suture suggest a key role for twist in suture initiation and biogenesis. *Mech Dev*, **2000**. 91(1-2): p. 341-5.

177. Rice, D.P., T. Aberg, Y. Chan, Z. Tang, P.J. Kettunen, L. Pakarinen, R.E. Maxson, and I. Thesleff, Integration of FGF and TWIST in calvarial bone and suture development. *Development*, **2000**. 127(9): p. 1845-55.

178. Price, J.A., D.W. Bowden, J.T. Wright, M.J. Pettenati, and T.C. Hart, Identification of a mutation in DLX3 associated with tricho-dento-osseous (TDO) syndrome. *Hum Mol Genet*, **1998**. 7(3): p. 563-9.

179. Uz, E., Y. Alanay, D. Aktas, I. Vargel, S. Gucer, G. Tuncbilek, F. von Eggeling, E. Yilmaz, *et al.*, Disruption of ALX1 causes extreme microphthalmia and severe facial clefting: expanding the spectrum of autosomal-recessive ALX-related frontonasal dysplasia. *Am J Hum Genet*, **2010**. 86(5): p. 789-96.

180. Ramos, C. and B. Robert, msh/Msx gene family in neural development. *Trends Genet*, **2005**. 21(11): p. 624-32.

181. Wuyts, W., W. Reardon, S. Preis, T. Homfray, A. Rasore-Quartino, H. Christians, P.J. Willems, and W. Van Hul, Identification of mutations in the MSX2 homeobox gene in families affected with foramina parietalia permagna. *Hum Mol Genet*, **2000**. 9(8): p. 1251-5.

182. Satokata, I., Msx2 deficiency in mice causes pleiotropic defects in bone growth and ectodermal organ formation. *Nature Genet.*, **2000**. 24: p. 391-395.

183. Ma, L., S. Golden, L. Wu, and R. Maxson, The molecular basis of Boston-type craniosynostosis: The Pro148->His mutation in the N-terminal arm of the MSX2 homeodomain stabilizes DNA binding without altering nucleotide sequence preferences. *Hum Mol Genet*, **1996**. 5(12): p. 1915-1920.

184. Jabs, E.W., U. Muller, X. Li, L. Ma, W. Luo, I.S. Haworth, I. Klisak, R. Sparkes, *et al.*, A mutation in the homeodomain of the human MSX2 gene in a family affected with autosomal dominant craniosynostosis. *Cell*, **1993**. 75(3): p. 443-50.

185. Nakatomi, M., X.P. Wang, D. Key, J.J. Lund, A. Turbe-Doan, R. Kist, A. Aw, Y. Chen, *et al.*, Genetic interactions between Pax9 and Msx1 regulate lip development and several stages of tooth morphogenesis. *Dev Biol*, **2010**. 340(2): p. 438-49.

186. Satokata, I. and R. Maas, Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development. *Nat Genet*, **1994**. 6(4): p. 348-56.

187. Van den Boogaard, M.J.H., M. Dorland, F.A. Beemer, and H.K. Amstel, MSX1 mutation is associated with orofacial clefting and tooth agenesis in humans. *Nat Genet*, **2000**. 24: p. 342-343.

188. Tzoulaki, I., I.M. White, and I.M. Hanson, PAX6 mutations: genotype-phenotype correlations. *BMC Genet*, **2005**. 6(1): p. 27.

189. Mathers, P.H. and M. Jamrich, Regulation of eye formation by the Rx and pax6 homeobox genes. *Cell Mol Life Sci*, **2000**. 57(2): p. 186-94.

190. Barrow, J.R. and M.R. Capecchi, Compensatory defects associated with mutations in Hoxa1 restore normal palatogenesis to Hoxa2 mutants. *Development*, **1999**. 126(22): p. 5011-26.

191. Mao, X.Y. and S.J. Tang, Effects of phenytoin on Satb2 and Hoxa2 gene expressions in mouse embryonic craniofacial tissue. *Biochem Cell Biol*, **2010**. 88(4): p. 731-5.

192. Bangs, F., M. Welten, M.G. Davey, M. Fisher, Y. Yin, H. Downie, B. Paton, R. Baldock, *et al.*, Identification of genes downstream of the Shh signalling in the developing chick wing and syn-expressed with Hoxd13 using microarray and 3D computational analysis. *Mech Dev*, **2010**. 127(9-12): p. 428-41.

193. Cooper, M.K., J.A. Porter, K.E. Young, and P.A. Beachy, Teratogen-mediated inhibition of target tissue response to Shh signaling. *Science*, **1998**. 280(5369): p. 1603-7.

194. Marigo, V., D.J. Roberts, S.M. Lee, O. Tsukurov, T. Levi, J.M. Gastier, D.J. Epstein, D.J. Gilbert, *et al.*, Cloning, expression, and chromosomal location of SHH and IHH: two human homologues of the Drosophila segment polarity gene hedgehog. *Genomics*, **1995**. 28(1): p. 44-51.

195. Paiva, K.B., M. Silva-Valenzuela, S.M. Massironi, G.M. Ko, F.M. Siqueira, and F.D. Nunes, Differential Shh, Bmp and Wnt gene expressions during craniofacial development in mice. *Acta Histochem*, **2010**. 112(5): p. 508-17.

196. Reid, B.S., H. Yang, V.S. Melvin, M.M. Taketo, and T. Williams, Ectodermal Wnt/beta-catenin signaling shapes the mouse face. *Dev Biol*, **2011**. 349(2): p. 261-9.

197. Young, N.M., H.J. Chong, D. Hu, B. Hallgrimsson, and R.S. Marcucio, Quantitative analyses link modulation of sonic hedgehog signaling to continuous variation in facial growth and shape. *Development*, **2010**. 137(20): p. 3405-9.

198. Liu, R., L. Wang, G. Chen, H. Katoh, C. Chen, Y. Liu, and P. Zheng, FOXP3 up-regulates p21 expression by site-specific inhibition of histone deacetylase

2/histone deacetylase 4 association to the locus. *Cancer Res*, **2009**. 69(6): p. 2252-9.

199. Liu, Y., Y. Wang, W. Li, P. Zheng, and Y. Liu, Activating transcription factor 2 and c-Jun-mediated induction of FoxP3 for experimental therapy of mammary tumor in the mouse. *Cancer Res*, **2009**. 69(14): p. 5954-60.

200. Hide, T., J. Hatakeyama, C. Kimura-Yoshida, E. Tian, N. Takeda, Y. Ushio, T. Shiroishi, S. Aizawa, *et al.*, Genetic modifiers of otocephalic phenotypes in Otx2 heterozygous mutant mice. *Development*, **2002**. 129(18): p. 4347-57.

201. Matsuo, I., S. Kuratani, C. Kimura, N. Takeda, and S. Aizawa, Mouse Otx2 functions in the formation and patterning of rostral head. *Genes Dev*, **1995**. 9(21): p. 2646-58.

202. Torres, M., E. Gomez-Pardo, and P. Gruss, Pax2 contributes to inner ear patterning and optic nerve trajectory. *Development*, **1996**. 122(11): p. 3381-91.

203. Jiang, R., Y. Lan, H.D. Chapman, C. Shawber, C.R. Norton, D.V. Serreze, G. Weinmaster, and T. Gridley, Defects in limb, craniofacial, and thymic development in Jagged2 mutant mice. *Genes & Development*, **1998**. 12(7): p. 1046-1057.

204. Zouvelou, V., H.U. Luder, T.A. Mitsiadis, and D. Graf, Deletion of BMP7 affects the development of bones, teeth, and other ectodermal appendages of the orofacial complex. *J Exp Zool B Mol Dev Evol*, **2009**. 312B(4): p. 361-74.

205. Li, Y., D.A. Lacerda, M.L. Warman, D.R. Beier, H. Yoshioka, Y. Ninomiya, J.T. Oxford, N.P. Morris, *et al.*, A Fibrillar Collagen Gene, Col11a1, Is Essential for Skeletal Morphogenesis. *Cell*, **1995**. 80(3): p. 423-430.

206. Vikkula, M., E.C. Mariman, V.C. Lui, N.I. Zhidkova, G.E. Tiller, M.B. Goldring, S.E. van Beersum, M.C. de Waal Malefijt, *et al.*, Autosomal dominant and recessive osteochondrodysplasias associated with the COL11A2 locus. *Cell*, **1995**. 80(3): p. 431-7.

207. Arikawa-Hirasawa, E., H. Watanabe, H. Takami, J.R. Hassell, and Y. Yamada, Perlecan is essential for cartilage and cephalic development. *Nature genetics*, **1999**. 23(3): p. 354-358.

208. Lavrin, I.O., W. McLean, R.E. Seegmiller, B.R. Olsen, and E.D. Hay, The mechanism of palatal clefting in the Col11a1 mutant mouse. *Arch Oral Biol*, **2001**. 46(9): p. 865-9.

209. Brugmann, S.A., K.E. Powder, N.M. Young, L.H. Goodnough, S.M. Hahn, A.W. James, J.A. Helms, and M. Lovett, Comparative gene expression analysis of avian embryonic facial structures reveals new candidates for human craniofacial disorders. *Hum Mol Genet*, **2010**. 19(5): p. 920-30.

210. Jiang, R., Y. Lan, H.D. Chapman, C. Shawber, C.R. Norton, D.V. Serreze, G. Weinmaster, and T. Gridley, Defects in limb, craniofacial, and thymic development in Jagged2 mutant mice. *Genes Dev*, **1998**. 12(7): p. 1046-57.

211. Zhang, Y., T. Mori, H. Takaki, M. Takeuch, K. Iseki, S. Hagino, M. Murakawa, S. Yokoya, *et al.*, Comparison of the expression patterns of two LIM-homeodomain genes, Lhx6 and L3/Lhx8, in the developing palate. *Orthod Craniofac Res*, **2002**. 5(2): p. 65-70.

212. Zhao, Y., Y.J. Guo, A.C. Tomac, N.R. Taylor, A. Grinberg, E.J. Lee, S. Huang, and H. Westphal, Isolated cleft palate in mice with a targeted mutation of the LIM homeobox gene lhx8. *Proc Natl Acad Sci U S A*, **1999**. 96(26): p. 15002-6.

213. Attanasio, C., A.S. Nord, Y. Zhu, M.J. Blow, Z. Li, D.K. Liberton, H. Morrison, I. Plajzer-Frick, *et al.*, Fine tuning of craniofacial morphology by distant-acting enhancers. *Science*, **2013**. 342(6157): p. 1241006.

214. Druzhkova, A.S., O. Thalmann, V.A. Trifonov, J.A. Leonard, N.V. Vorobieva, N.D. Ovodov, A.S. Graphodatsky, and R.K. Wayne, Ancient DNA analysis affirms the canid from Altai as a primitive dog. *PLoS One*, **2013**. 8(3): p. e57754.

215. Vila, C., P. Savolainen, J.E. Maldonado, I.R. Amorim, J.E. Rice, R.L. Honeycutt, K.A. Crandall, J. Lundeberg, *et al.*, Multiple and ancient origins of the domestic dog. *Science*, **1997**. 276(5319): p. 1687-9.

216. Freedman, A.H., I. Gronau, R.M. Schweizer, D. Ortega-Del Vecchyo, E. Han, P.M. Silva, M. Galaverni, Z. Fan, *et al.*, Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet*, **2014**. 10(1): p. e1004016.

217. Karlsson, E.K., I. Baranowska, C.M. Wade, N.H. Salmon Hillbertz, M.C. Zody, N. Anderson, T.M. Biagi, N. Patterson, *et al.*, Efficient mapping of mendelian traits in dogs through genome-wide association. *Nat Genet*, **2007**. 39(11): p. 1321-8.

218. Schoenebeck, J.J. and E.A. Ostrander, The genetics of canine skull shape variation. *Genetics*, **2013**. 193(2): p. 317-25.

219. Kerstin Lindblad-Toh1, C.M.W., 2, Tarjei S. Mikkelsen1,3, Elinor K. Karlsson1,4, David B. Jaffe1,, M.C. Michael Kamal1, Jean L. Chang1, Edward J. Kulbokas III1, Michael C. Zody1, Evan Mauceli1,, M.B. Xiaohui Xie1, Robert K. Wayne6, Elaine A. Ostrander7, Chris P. Ponting8, Francis Galibert9,, P.J.d. Douglas R. Smith10, Ewen Kirkness12, Pablo Alvarez1, Tara Biagi1, William Brockman1,, C.-W.C. Jonathan Butler1, April Cook1, James Cuff1, Mark J. Daly1,2, David DeCaprio1, Sante Gnerre1,, M.K. Manfred Grabherr1, 13, Michael Kleber1, Carolyne Bardeleben6, Leo Goodstadt8, Andreas Heger8,, L.K. Christophe Hitte9, Klaus-Peter Koepfli6, Heidi G. Parker7, John P. Pollinger6, Stephen M. J. Searle14,, and R.T. Nathan B. Sutter7, Caleb Webber8, Broad Institute Genome Sequencing Platform, Eric S. Lander, Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **2005**. 438.

220. Barsh, G.S., How the dog got its spots. *Nat Genet*, **2007**. 39(11): p. 1304-6.

221. Schoenebeck, J.J., S.A. Hutchinson, A. Byers, H.C. Beale, B. Carrington, D.L. Faden, M. Rimbault, B. Decker, *et al.*, Variation of BMP3 contributes to dog breed skull diversity. *PLoS Genet*, **2012**. 8(8): p. e1002849.

222. Bannasch, D., A. Young, J. Myers, K. Truve, P. Dickinson, J. Gregg, R. Davis, E. Bongcam-Rudloff, *et al.*, Localization of canine brachycephaly using an across breed mapping approach. *PLoS One*, **2010**. 5(3): p. e9632.

223. Quilez, J., A.D. Short, V. Martinez, L.J. Kennedy, W. Ollier, A. Sanchez, L. Altet, and O. Francino, A selective sweep of >8 Mb on chromosome 26 in the Boxer genome. *BMC Genomics*, **2011**. 12: p. 339.

224. Lindblad-Toh, K., C.M. Wade, T.S. Mikkelsen, E.K. Karlsson, D.B. Jaffe, M. Kamal, M. Clamp, J.L. Chang, *et al.*, Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **2005**. 438(7069): p. 803-819.

225. Karlssons, E.K. and K. Linblad-Toh, Leader of the pack: gene mapping in dogs and other model organisms. *Nature Rev. Genet.*, **2008**. 9: p. 713.

226. Akey, J.M., A.L. Ruhe, D.T. Akey, A.K. Wong, C.F. Connelly, J. Madeoy, T.J. Nicholas, and M.W. Neff, Tracking footprints of artificial selection in the dog genome. *Proc Natl Acad Sci U S A*, **2010**. 107(3): p. 1160-5.

227. Sherwood, R.J., D.L. Duren, L.M. Havill, J. Rogers, L.A. Cox, B. Towne, and M.C. Mahaney, A genomewide linkage scan for quantitative trait loci influencing the craniofacial complex in baboons (Papio hamadryas spp.). *Genetics*, **2008**. 180(1): p. 619-28.

228. Sherwood, R.J., D.L. Duren, M.C. Mahaney, J. Blangero, T.D. Dyer, S.A. Cole, S.A. Czerwinski, W.C. Chumlea, *et al.*, A genome-wide linkage scan for quantitative trait loci influencing the craniofacial complex in humans (Homo sapiens sapiens). *Anat Rec (Hoboken)*, **2011**. 294(4): p. 664-75.

229. Skipper, M., Human genetics - Not-so-identical twins. *Nat Rev Genet*, **2008**. 9(4): p. 249-249.

230. Roth, S.Y., J.M. Denu, and C.D. Allis, Histone acetyltransferases. *Annu Rev Biochem*, **2001**. 70(1): p. 81-120.

231. Ptashne, M., On the use of the word 'epigenetic'. *Curr Biol*, **2007**. 17(7): p. R233-6.

232. Clayton, A.L., C.A. Hazzalin, and L.C. Mahadevan, Enhanced histone acetylation and transcription: a dynamic perspective. *Mol Cell*, **2006**. 23(3): p. 289-96.

233. Volpe, T.A., C. Kidner, I.M. Hall, G. Teng, S.I. Grewal, and R.A. Martienssen, Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, **2002**. 297(5588): p. 1833-7.

234. Strahl, B.D. and C.D. Allis, The language of covalent histone modifications. *Nature*, **2000**. 403(6765): p. 41-5.

235. Jenuwein, T. and C.D. Allis, Translating the histone code. *Science*, **2001**. 293(5532): p. 1074-80.

236. Wang, Z., C. Zang, J.A. Rosenfeld, D.E. Schones, A. Barski, S. Cuddapah, K. Cui, T.Y. Roh, *et al.*, Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat Genet*, **2008**. 40(7): p. 897-903.

237. Kidd, J.M., G.M. Cooper, W.F. Donahue, H.S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, *et al.*, Mapping and sequencing of structural variation from eight human genomes. *Nature*, **2008**. 453(7191): p. 56-64.

238. Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, *et al.*, Global variation in copy number in the human genome. *Nature*, **2006**. 444(7118): p. 444-54.

239. Montgomery, M.K., S.Q. Xu, and A. Fire, RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans. *Proc Natl Acad Sci U S A*, **1998**. 95(26): p. 15502-15507.

240. Hannon, G.J. and J.J. Rossi, Unlocking the potential of the human genome with RNA interference. *Nature*, **2004**. 431(7006): p. 371-8.

241. Mattick, J.S., P.P. Amaral, M.E. Dinger, T.R. Mercer, and M.F. Mehler, RNA regulation of epigenetic processes. *Bioessays*, **2009**. 31(1): p. 51-9.

242. Hamilton, A.J. and D.C. Baulcombe, A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, **1999**. 286(5441): p. 950-2.

243. Mattick, J.S. and I.V. Makunin, Non-coding RNA. *Hum Mol Genet*, **2006**. 15 Spec No 1(suppl 1): p. R17-29.

244. Sherry, S.T., M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, and K. Sirotkin, dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **2001**. 29(1): p. 308-11.

245. Vapnik, V., The International HapMap Consortium. The international Hapmap Project. *Nature*, **2003**. 426: p. 789 - 796.

246. Genomes Project, C., G.R. Abecasis, D. Altshuler, A. Auton, L.D. Brooks, R.M. Durbin, R.A. Gibbs, M.E. Hurles, *et al.*, A map of human genome variation from population-scale sequencing. *Nature*, **2010**. 467(7319): p. 1061-73.

247. Budowle, B. and A. van Daal, Forensically relevant SNP classes. *Biotechniques*, **2008**. 44(5): p. 603-8, 610.

248. Valenzuela, R.K., M.S. Henderson, M.H. Walsh, N.A. Garrison, J.T. Kelch, O. Cohen-Barak, D.T. Erickson, F. John Meaney, *et al.*, Predicting Phenotype from Genotype: Normal Pigmentation*. *Journal of Forensic Sciences*. 55(2): p. 315-322.

249. Frudakis, T., K. Venkateswarlu, M.J. Thomas, Z. Gaskin, S. Ginjupalli, S. Gunturi, V. Ponnuswamy, S. Natarajan, *et al.*, A classifier for the SNP-based inference of ancestry. *J Forensic Sci*, **2003**. 48(4): p. 771-82.

250. Izagirre, N., I. Garcia, C. Junquera, C. de la Rua, and S. Alonso, A scan for signatures of positive selection in candidate loci for skin pigmentation in humans. *Mol Biol Evol*, **2006**. 23(9): p. 1697-706.

251. Lao, O., J.M. de Gruijter, K. van Duijn, A. Navarro, and M. Kayser, Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann Hum Genet*, **2007**. 71(Pt 3): p. 354-69.

252. Pneuman, A., Z.M. Budimlija, T. Caragine, M. Prinz, and E. Wurmbach, Verification of eye and skin color predictors in various populations. *Legal Medicine*, **2012**. 14(2): p. 78-83.

253. Biswas, S. and J.M. Akey, Genomic insights into positive selection. *Trends Genet*, **2006**. 22(8): p. 437-46.

254. Huff, C.D., H.C. Harpending, and A.R. Rogers, Detecting positive selection from genome scans of linkage disequilibrium. *BMC Genomics*, **2010**. 11(1): p. 8.

255. Sabeti, P.C., D.E. Reich, J.M. Higgins, H.Z. Levine, D.J. Richter, S.F. Schaffner, S.B. Gabriel, J.V. Platko, *et al.*, Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **2002**. 419(6909): p. 832-7.

256. Voight, B.F., S. Kudaravalli, X. Wen, and J.K. Pritchard, A map of recent positive selection in the human genome. *PLoS Biol*, **2006**. 4: p. e72.

257. Barreiro, L.B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, Natural selection has driven population differentiation in modern humans. *Nat Genet*, **2008**. 40(3): p. 340-5.

258. Sabeti, P.C., S.F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T.S. Mikkelsen, *et al.*, Positive natural selection in the human lineage. *Science*, **2006**. 312(5780): p. 1614-20.

259. Sabeti, P.C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E.H. Byrne, *et al.*, Genome-wide detection and characterization of positive selection in human populations. *Nature*, **2007**. 449: p. 913 - 918.

260. Coop, G., J.K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R.M. Myers, L.L. Cavalli-Sforza, *et al.*, The role of geography in human adaptation. *PLoS Genet*, **2009**. 5(6): p. e1000500.

261. Akey, J.M., G. Zhang, K. Zhang, L. Jin, and M.D. Shriver, Interrogating a high-density SNP map for signatures of natural selection. *Genome Res*, **2002**. 12(12): p. 1805-14.

262. Pickrell, J.K., G. Coop, J. Novembre, S. Kudaravalli, J.Z. Li, D. Absher, B.S. Srinivasan, G.S. Barsh, *et al.*, Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*, **2009**. 19(5): p. 826-37.

263. Przeworski, M., G. Coop, and J.D. Wall, The signature of positive selection on standing genetic variation. *Evolution*, **2005**. 59(11): p. 2312-23.

264. Eriksson, N., J.M. Macpherson, J.Y. Tung, L.S. Hon, B. Naughton, S. Saxonov, L. Avey, A. Wojcicki, *et al.*, Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet*, **2010**. 6(6): p. e1000993.

265. Paschou, P., J. Lewis, A. Javed, and P. Drineas, Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J Med Genet*, **2010**. 47(12): p. 835-47.

266. Li, J.Z., D.M. Absher, H. Tang, A.M. Southwick, A.M. Casto, S. Ramachandran, H.M. Cann, G.S. Barsh, *et al.*, Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **2008**. 319(5866): p. 1100-4.

267. Halder, I., M. Shriver, M. Thomas, J.R. Fernandez, and T. Frudakis, A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Hum Mutat*, **2008**. 29(5): p. 648-58.

268. Londin, E.R., M.A. Keller, C. Maista, G. Smith, L.A. Mamounas, R. Zhang, S.J. Madore, K. Gwinn, *et al.*, CoAIMs: a cost-effective panel of ancestry informative markers for determining continental origins. *PLoS One*, **2010**. 5(10): p. e13443.

269. Capon, F., M.H. Allen, M. Ameen, A.D. Burden, D. Tillman, J.N. Barker, and R.C. Trembath, A synonymous SNP of the corneodesmosin gene leads to increased mRNA stability and demonstrates association with psoriasis across diverse ethnic groups. *Hum Mol Genet*, **2004**. 13(20): p. 2361-8.

270. Sauna, Z.E., C. Kimchi-Sarfaty, S.V. Ambudkar, and M.M. Gottesman, Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer. *Cancer Res*, **2007**. 67(20): p. 9609-12.

271. Duan, J., M.S. Wainwright, J.M. Comeron, N. Saitou, A.R. Sanders, J. Gelernter, and P.V. Gejman, Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Hum Mol Genet*, **2003**. 12(3): p. 205-16.

272. Hunt, R., Z.E. Sauna, S.V. Ambudkar, M.M. Gottesman, and C. Kimchi-Sarfaty, *Silent (synonymous) SNPs: should we care about them?*, in Methods in Molecular Biology. **2009**, Springer. p. 23-39.

273. Sauna, Z.E. and C. Kimchi-Sarfaty, Understanding the contribution of synonymous mutations to human disease. *Nat Rev Genet*, **2011**. 12(10): p. 683-91.

274. Wang, E.T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S.F. Kingsmore, G.P. Schroth, *et al.*, Alternative isoform regulation in human tissue transcriptomes. *Nature*, **2008**. 456(7221): p. 470-6.

275. Wang, Z. and C.B. Burge, Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **2008**. 14(5): p. 802-13.

276. Chen, K. and N. Rajewsky, The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*, **2007**. 8(2): p. 93-103.

277. The International HapMap, C., A haplotype map of the human genome. *Nature*, **2005**. 437(7063): p. 1299 - 1320.

278. Butler, J.M., M.D. Coble, and P.M. Vallone, STRs vs. SNPs: thoughts on the future of forensic DNA testing. *Forensic Sci Med Pathol*, **2007**. 3(3): p. 200-205.

279. Butler, J.M., Y. Shen, and B.R. McCord, The development of reduced size STR amplicons as tools for analysis of degraded DNA. *J Forensic Sci*, **2003**. 48(5): p. 1054-1064.

280. Foster, E.A., M.A. Jobling, P.G. Taylor, P. Donnelly, P. de Knijff, R. Mieremet, T. Zerjal, and C. Tyler-Smith, Jefferson fathered slave's last child. *Nature*, **1998**. 396(6706): p. 27-8.

281. Budowle, B., F.R. Bieber, and A.J. Eisenberg, Forensic aspects of mass disasters: strategic considerations for DNA-based human identification. *Leg Med (Tokyo)*, **2005**. 7(4): p. 230-43.

282. Coble, M.D., O.M. Loreille, M.J. Wadhams, S.M. Edson, K. Maynard, C.E. Meyer, H. Niederstatter, C. Berger, *et al.*, Mystery Solved: The Identification of the Two Missing Romanov Children Using DNA Analysis. *Plos One*, **2009**. 4(3): p. e4838.

283. Musgrave-Brown, E., D. Ballard, K. Balogh, K. Bender, B. Berger, M. Bogus, C. Borsting, M. Brion, *et al.*, Forensic validation of the SNPforID 52-plex assay. *Forensic Sci Int Genet*, **2007**. 1(2): p. 186-90.

284. Pakstis, A.J., W.C. Speed, J.R. Kidd, and K.K. Kidd, Candidate SNPs for a universal individual identification panel. *Hum Genet*, **2007**. 121(3-4): p. 305-17.

285. Tobler, A.R., S. Short, M.R. Andersen, T.M. Paner, J.C. Briggs, S.M. Lambert, P.P. Wu, Y. Wang, *et al.*, The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. *J Biomol Tech*, **2005**. 16(4): p. 398-406.

286. Rixun, F., J.P. Andrew, H. Fiona, W. David, S. Jaiprakash, R.K. Judith, K.K. Kenneth, and R.F. Manohar, Multiplexed SNP detection panels for human identification. *Forensic Sci Inter: Genet Suppl*, **2009**. 2(1): p. 538-539.

287. Sanchez, J.J., C. Phillips, C. Borsting, K. Balogh, M. Bogus, M. Fondevila, C.D. Harrison, E. Musgrave-Brown, *et al.*, A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis*, **2006**. 27(9): p. 1713-24.

288. Pakstis, A.J., W.C. Speed, R. Fang, F.C. Hyland, M.R. Furtado, J.R. Kidd, and K.K. Kidd, SNPs for a universal individual identification panel. *Hum Genet*, **2010**. 127(3): p. 315-24.

289. Kidd, K.K., A.J. Pakstis, W.C. Speed, E.L. Grigorenko, S.L. Kajuna, N.J. Karoma, S. Kungulilo, J.J. Kim, *et al.*, Developing a SNP panel for forensic identification of individuals. *Forensic Sci Int*, **2006**. 164(1): p. 20-32.

290. Ge, J., B. Budowle, J.V. Planz, and R. Chakraborty, Haplotype block: a new type of forensic DNA markers. *Int J Legal Med*, **2010**. 124(5): p. 353-61.

291. Phillips, C., A. Salas, J.J. Sanchez, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, M. Calaza, M.C. de Cal, *et al.*, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet*, **2007**. 1(3-4): p. 273-80.

292. Walsh, S., A. Lindenbergh, S.B. Zuniga, T. Sijen, P. de Knijff, M. Kayser, and K.N. Ballantyne, Developmental validation of the IrisPlex system: determination of blue and brown iris colour for forensic intelligence. *Forensic Sci Int Genet*, **2011**. 5(5): p. 464-71.

293. Walsh, S., F. Liu, K.N. Ballantyne, M. van Oven, O. Lao, and M. Kayser, IrisPlex: a sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci Int Genet*, **2011**. 5(3): p. 170-80.

294. Grimes, E.A., P.J. Noake, L. Dixon, and A. Urquhart, Sequence polymorphism in the human melanocortin 1 receptor gene as an indicator of the red hair phenotype. *Forensic Science International*, **2001**. 122(2-3): p. 124-129.

295. Mengel-From, J., C. Borsting, J.J. Sanchez, H. Eiberg, and N. Morling, Human eye colour and HERC2, OCA2 and MATP. *Forensic Sci Int Genet*, **2010**. 4(5): p. 323-8.

296. Walsh, S., F. Liu, A. Wollstein, L. Kovatsi, A. Ralf, A. Kosiniak-Kamysz, W. Branicki, and M. Kayser, The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Sci Int Genet*, **2013**. 7(1): p. 98-115.

297. Walsh, S., A. Wollstein, F. Liu, U. Chakravarthy, M. Rahu, J.H. Seland, G. Soubrane, L. Tomazzoli, *et al.*, DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Sci Int Genet*, **2012**. 6(3): p. 330-40.

298. Weedon, M.N., H. Lango, C.M. Lindgren, C. Wallace, D.M. Evans, M. Mangino, R.M. Freathy, J.R. Perry, *et al.*, Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet*, **2008**. 40(5): p. 575-83.

299. Lango Allen, H., K. Estrada, G. Lettre, S.I. Berndt, M.N. Weedon, F. Rivadeneira, C.J. Willer, A.U. Jackson, *et al.*, Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **2010**. 467(7317): p. 832-8.

300. Lettre, G., A.U. Jackson, C. Gieger, F.R. Schumacher, S.I. Berndt, S. Sanna, S. Eyheramendy, B.F. Voight, *et al.*, Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet*, **2008**. 40(5): p. 584-91.

301. Gill, P., An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int J Legal Med*, **2001**. 114(4-5): p. 204-10.

302. Phillips, C., M. Lareu, J. Sanchez, M. Brion, B. Sobrino, N. Morling, P. Schneider, D.S. Court, *et al.*, Selecting single nucleotide polymorphisms for forensic applications. *Prog Foren Genet*, **2004**. 1261(10): p. 18-20.

303. Butler, J.M., B. Budowle, P. Gill, K.K. Kidd, C. Phillips, P.M. Schneider, P.M. Vallone, and N. Morling, Report on ISFG SNP Panel Discussion. *Forensic Sci Inter Genet Supp*, **2008**. 1(1): p. 471-472.

304. Sobrino, B., M. Brion, and A. Carracedo, SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic Sci Int*, **2005**. 154(2-3): p. 181-94.

305. Shen, R., J.B. Fan, D. Campbell, W. Chang, J. Chen, D. Doucet, J. Yeakley, M. Bibikova, *et al.*, High-throughput SNP genotyping on universal bead arrays. *Mutat Res*, **2005**. 573(1-2): p. 70-82.

306. Dixon, L.A., A.E. Dobbins, H.K. Pulker, J.M. Butler, P.M. Vallone, M.D. Coble, W. Parson, B. Berger, *et al.*, Analysis of artificially degraded DNA using STRs and SNPs--results of a collaborative European (EDNAP) exercise. *Forensic Sci Int*, **2006**. 164(1): p. 33-44.

307. Sanchez, J.J., C. Borsting, K. Balogh, B. Berger, M. Bogus, J.M. Butler, A. Carracedo, D.S. Court, *et al.*, Forensic typing of autosomal SNPs with a 29 SNP-multiplex--results of a collaborative EDNAP exercise. *Forensic Sci Int Genet*, **2008**. 2(3): p. 176-83.

308. Coassin, S., A. Brandstatter, and F. Kronenberg, Lost in the space of bioinformatic tools: a constantly updated survival guide for genetic epidemiology. The GenEpi Toolbox. *Atherosclerosis*, **2010**. 209(2): p. 321-35.

309. Kwok, P.Y., Methods for genotyping single nucleotide polymorphisms. *Annu Rev Genomics Hum Genet*, **2001**. 2(1): p. 235-58.

310. Berglund, E.C., A. Kiialainen, and A.C. Syvanen, Next-generation sequencing technologies and applications for human genetic history and forensics. *Investig Genet*, **2011**. 2(1): p. 23.

311. Shendure, J. and H. Ji, Next-generation DNA sequencing. *Nat Biotechnol*, **2008**. 26(10): p. 1135-45.

312. Hert, D.G., C.P. Fredlake, and A.E. Barron, Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, **2008**. 29(23): p. 4618-26.

313. Van Neste, C., F. Van Nieuwerburgh, D. Van Hoofstat, and D. Deforce, Forensic STR analysis using massive parallel sequencing. *Forensic Sci Int Genet*, **2012**. 6(6): p. 810-818.

314. Gymrek, M., D. Golan, S. Rosset, and Y. Erlich, lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res*, **2012**. 22(6): p. 1154-62.

315. Warshauer, D.H., D. Lin, K. Hari, R. Jain, C. Davis, B. Larue, J.L. King, and B. Budowle, STRait Razor: a length-based forensic STR allele-calling tool for use with second generation sequencing data. *Forensic Sci Int Genet*, **2013**. 7(4): p. 409-17.

316. Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J.V. Pearson, D.A. Stephan, *et al.*, Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genet*, **2008**. 4(8): p. e1000167.

317. Planz, J.V., B. Budowle, T. Hall, A.J. Eisenberg, K.A. Sannes-Lowery, and S.A. Hofstadler, Enhancing resolution and statistical power by utilizing mass

spectrometry for detection of SNPs within the short tandem repeats. *Forensic Sci Int Genet Suppl*, **2009**. 2(1): p. 529-531.

318. Quail, M.A., M. Smith, P. Coupland, T.D. Otto, S.R. Harris, T.R. Connor, A. Bertoni, H.P. Swerdlow, *et al.*, A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **2012**. 13(1): p. 341.

319. Rothberg, J.M., W. Hinz, T.M. Rearick, J. Schultz, W. Mileski, M. Davey, J.H. Leamon, K. Johnson, *et al.*, An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **2011**. 475(7356): p. 348-52.

320. Solutions, M.B. Instructions for Isohelix DNA Isolation kits: DDK-3/DDK-50. **June 2012**.

321. Qiagen. QIAprep Miniprep. **2014 accessed on 12/08/2014** [cited 2014 12 August]; Available from: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CCkQFjAA&url=http%3A%2F%2Fwww.qiagen.com%2Fresources%2Fdownload.aspx%3Fid%3D89bfa021-7310-4c0f-90e0-6a9c84f66cee%26lang%3Den&ei=B8LpU_ykIIbk8AWH4IKgCA&usg=AFQjCNFJmDtRJvlQgMZ4h5-bFyevLKmQUA&sig2=Qt1sB7SOtABhz4sqPGRGJA&bvm=bv.72676100,d.dGc&cad=rja.

322. Promega. Promega DNA IQ protocol. **2014 12/08/2014** [cited 2014 12 August]; Available from: http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=3&cad=rja&uact=8&ved=0CDQQFjAC&url=http%3A%2F%2Fwww.promega.com%2F~%2Fmedia%2Ffiles%2Fresources%2Fprotocols%2Ftechnical%2520bulletins%2F101%2Fdna%2520iq%2520system%2520database%2520protocol.pdf&ei=TMPpU-O0DZPh8AXBloHAAg&usg=AFQjCNE2dHsI8CDq_nKrmRM0hN0H0-y9SA&sig2=puK0te6QdrJVXrEjJVubNw&bvm=bv.72676100,d.dGc.

323. Rebhan, M., V. ChalifaCaspi, J. Prilusky, and D. Lancet, GeneCards: Integrating information about genes, proteins and diseases. *Trends in Genetics*, **1997**. 13(4): p. 163-163.

324. Harris, M.A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, *et al.*, The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, **2004**. 32(Database issue): p. D258-61.

325. Amigo, J., A. Salas, and C. Phillips, ENGINES: exploring single nucleotide variation in entire human genomes. *BMC Bioinformatics*, **2011**. 12(1): p. 105.

326. Shiwei Duan, Wei Zhang, N.J. Cox, and M.E. Dolan, FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3. *Bioinformation*, **2008**. 3(3): p. 139-141.

327. Sabeti, P.C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E.H. Byrne, *et al.*, Genome-wide detection and characterization of positive selection in human populations. *Nature*, **2007**. 449(7164): p. 913-8.

328. Cheng, F., W. Chen, E. Richards, L. Deng, and C. Zeng, SNP@Evolution: a hierarchical database of positive selection on the human genome. *BMC Evol Biol*, **2009**. 9(1): p. 221.

329. Amigo, J., A. Salas, C. Phillips, and A. Carracedo, SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics*, **2008**. 9: p. 428.

330. Ramensky, V., Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*, **2002**. 30(17): p. 3894-3900.

331. Macintyre, G., J. Bailey, I. Haviv, and A. Kowalczyk, is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, **2010**. 26(18): p. i524-30.

332. Chelala, C., A. Khan, and N.R. Lemoine, SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, **2009**. 25(5): p. 655-61.

333. Wang, J., M. Ronaghi, S.S. Chong, and C.G. Lee, pfSNP: An integrated potentially functional SNP resource that facilitates hypotheses generation through knowledge syntheses. *Hum Mutat*, **2011**. 32(1): p. 19-24.

334. Barrett, J.C., B. Fry, J. Maller, and M.J. Daly, Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **2005**. 21(2): p. 263-5.

335. Edlund, C.K., W.H. Lee, D. Li, D.J. Van Den Berg, and D.V. Conti, Snagger: a user-friendly program for incorporating additional information for tagSNP selection. *BMC Bioinformatics*, **2008**. 9(1): p. 174.

336. Pettersson, F.H., C.A. Anderson, G.M. Clarke, J.C. Barrett, L.R. Cardon, A.P. Morris, and K.T. Zondervan, Marker selection for genetic case-control association studies. *Nat Protoc*, **2009**. 4(5): p. 743-52.

337. Carlson, C.S., M.A. Eberle, M.J. Rieder, Q. Yi, L. Kruglyak, and D.A. Nickerson, Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet*, **2004**. 74(1): p. 106-20.

338. Sicotte, H., D.N. Rider, G.A. Poland, N. Dhiman, and J.P. Kocher, SNPPicker: high quality tag SNP selection across multiple populations. *BMC Bioinformatics*, **2011**. 12(1): p. 129.

339. Swennen, G.J., F.A. Schuytser, and J.E. Hausamen, *Three dimensional cephalometry: a color atlas and manual* Vol. 28. **2006**: Springer. 195 p.

340. Technologies, L. Ion AmpliSeq™ DNA and RNA Library Preparation. **2014** [cited 2014 13 August]; Available from: http://ioncommunity.lifetechnologies.com/docs/DOC-3005.

341. Agencourt. AMPure XP protocol. **2014** [cited 2014 13 August]; Available from: https://www.beckmancoulter.com/wsrportal/bibliography?docname=Protocol_000387v001.pdf.

342. Technologies, L. Ion PGM™ Template OT2 200 Kit v2. **2014** [cited 2014 13 August]; Available from:

http://hwwgenotyping.ksu.edu/protocols/MAN0007624_Ion%20PI%20Template%20OT2%20200%20Kit%20v2_UG_Rev2_01July2013.pdf.

343. Technologies, L. Ion PGM™ Sequencing 200 Kit v2. **2014** [cited 2014 13 August]; Available from: http://download.bioon.com.cn/view/upload/201312/14155016_8616.pdf.

344. Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, *et al.*, PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, **2007**. 81(3): p. 559-75.

345. Gordon, D., S.J. Finch, M. Nothnagel, and J. Ott, Power and sample size calculations for case-control genetic association tests when errors are present: Application to single nucleotide polymorphisms. *Human Heredity*, **2002**. 54(1): p. 22-33.

346. Pfeiffer, R.M. and M.H. Gail, Sample size calculations for population- and family-based case-control association studies on marker genotypes. *Genet Epidemiol*, **2003**. 25(2): p. 136-48.

347. Laurie, C.C., K.F. Doheny, D.B. Mirel, E.W. Pugh, L.J. Bierut, T. Bhangale, F. Boehm, N.E. Caporaso, *et al.*, Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol*, **2010**. 34(6): p. 591-602.

348. Hong, E.P. and J.W. Park, Sample size and statistical power calculation in genetic association studies. *Genomics Inform*, **2012**. 10(2): p. 117-22.

349. Hughes-Stamm, S., M. Barash, K. Grisedale, and A. van Daal, Initial Evaluation of A96-Plex Goldengate® Genotyping SNP Assay with Suboptimal and Whole Genome Amplified Samples. *J Forensic Investigation*, **2013**. 1(1): p. 8.

350. Hughes-Stamm, S.R., DNA Typing Methods for Highly Degraded Samples, in Faculty of Health Sciences and Medicine. **2012**, *Bond University*: ePublications@bond. p. 419.

351. Wamalwa, P., S.K. Amisi, Y. Wang, and S. Chen, Angular photogrammetric comparison of the soft-tissue facial profile of Kenyans and Chinese. *J Craniofac Surg*, **2011**. 22(3): p. 1064-72.

352. Farkas, L.G., W. Bryson, and J. Klotz, Is photogrammetry of the face reliable? *Plast Reconstr Surg*, **1980**. 66(3): p. 346-355.

353. Nechala, P., J. Mahoney, and L.G. Farkas, Digital two-dimensional photogrammetry: a comparison of three techniques of obtaining digital photographs. *Plast Reconstr Surg*, **1999**. 103(7): p. 1819-25.

354. Kovacs, L., A. Zimmermann, G. Brockmann, H. Baurecht, K. Schwenzer-Zimmerer, N.A. Papadopulos, M.A. Papadopoulos, R. Sader, *et al.*, Accuracy and Precision of the Three-Dimensional Assessment of the Facial Surface Using a 3-D Laser Scanner. *IEEE Transactions on Medical Imaging*, **2006**. 25(6): p. 742-754.

355. Fourie, Z., J. Damstra, P.O. Gerrits, and Y. Ren, Evaluation of anthropometric accuracy and reliability using different three-dimensional scanning systems. *Forensic Sci Int*, **2011**. 207(1-3): p. 127-34.

356. Bianchi, S.D., M.C. Spada, L. Bianchi, L. VerzÃ¨, E. Vezzetti, S. Tornincasa, and G. Ramieri, Evaluation of scanning parameters for a surface colour laser scanner. *International Congress Series*, **2004**. 1268: p. 1162-1167.

357. Aung, S.C., R.C. Ngim, and S.T. Lee, Evaluation of the laser scanner as a surface measuring tool and its accuracy compared with direct facial anthropometric measurements. *Br J Plast Surg*, **1995**. 48(8): p. 551-8.

358. Kau, C.H., S. Richmond, A.I. Zhurov, J. Knox, I. Chestnutt, F. Hartles, and R. Playle, Reliability of measuring facial morphology with a 3-dimensional laser scanning system. *Am J Orthod Dentofacial Orthop*, **2005**. 128(4): p. 424-30.

359. Ma, L., T. Xu, and J. Lin, Validation of a three-dimensional facial scanning system based on structured light techniques. *Comput Methods Programs Biomed*, **2009**. 94(3): p. 290-8.

360. Plooij, J.M., G.R. Swennen, F.A. Rangel, T.J. Maal, F.A. Schutyser, E.M. Bronkhorst, A.M. Kuijpers-Jagtman, and S.J. Berge, Evaluation of reproducibility and reliability of 3D soft tissue analysis using 3D stereophotogrammetry. *Int J Oral Maxillofac Surg*, **2009**. 38(3): p. 267-73.

361. Shibata, M., H. Nawa, Y. Kise, M. Fuyamada, K. Yoshida, A. Katsumata, E. Ariji, and S. Goto, Reproducibility of three-dimensional coordinate systems based on craniofacial landmarks: a tentative evaluation of four systems created on images obtained by cone-beam computed tomography with a large field of view. *Angle Orthod*, **2012**. 82(5): p. 776-84.

362. Wong, J.Y., A.K. Oh, E. Ohta, A.T. Hunt, G.F. Rogers, J.B. Mulliken, and C.K. Deutsch, Validity and reliability of craniofacial anthropometric measurement of 3D digital photogrammetric images. *Cleft Palate Craniofac J*, **2008**. 45(3): p. 232-9.

363. Gwilliam, J.R., S.J. Cunningham, and T. Hutton, Reproducibility of soft tissue landmarks on three-dimensional facial scans. *Eur J Orthod*, **2006**. 28(5): p. 408-15.

364. Royston, P., Estimating departure from normality. *Stat Med*, **1991**. 10(8): p. 1283-93.

365. Altman, D.G. and J.M. Bland, Statistics notes: the normal distribution. *BMJ*, **1995**. 310(6975): p. 298.

366. Micceri, T., The Unicorn, the Normal Curve, and Other Improbable Creatures. *Psychological Bulletin*, **1989**. 105(1): p. 156-166.

367. O'Boyle, E. and H. Aguinis, The Best and the Rest: Revisiting the Norm of Normality of Individual Performance. *Personnel Psychology*, **2012**. 65(1): p. 79-119.

368. Royston, J., An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, **1982**: p. 115-124.

369. Field, A., *Discovering statistics using IBM SPSS statistics*. **2013**: Sage. 915 p.

370. Olalde, I., M.E. Allentoft, F. Sanchez-Quinto, G. Santpere, C.W. Chiang, M. DeGiorgio, J. Prado-Martinez, J.A. Rodriguez, *et al.*, Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, **2014**. 507(7491): p. 225-8.

371. Parra, E.J., R.A. Kittles, and M.D. Shriver, Implications of correlations between skin color and genetic ancestry for biomedical research. *Nat Genet*, **2004**. 36(11 Suppl): p. S54-60.

372. Weiss, K.M. and J.C. Long, Non-Darwinian estimation: my ancestors, my genes' ancestors. *Genome Res*, **2009**. 19(5): p. 703-10.

373. Shriver, M.D., E.J. Parra, S. Dios, C. Bonilla, H. Norton, C. Jovel, C. Pfaff, C. Jones, *et al.*, Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum Genet*, **2003**. 112(4): p. 387-99.

374. Rosenberg, N.A., L.M. Li, R. Ward, and J.K. Pritchard, Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet*, **2003**. 73(6): p. 1402-22.

375. Gettings, K.B., R. Lai, J.L. Johnson, M.A. Peck, J.A. Hart, H. Gordish-Dressman, M.S. Schanfield, and D.S. Podini, A 50-SNP assay for biogeographic ancestry and phenotype prediction in the U.S. population. *Forensic Sci Int Genet*, **2014**. 8(1): p. 101-8.

376. Kosoy, R., R. Nassir, C. Tian, P.A. White, L.M. Butler, G. Silva, R. Kittles, M.E. Alarcon-Riquelme, *et al.*, Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat*, **2009**. 30(1): p. 69-78.

377. Keating, B., A.T. Bansal, S. Walsh, J. Millman, J. Newman, K. Kidd, B. Budowle, A. Eisenberg, *et al.*, First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. *Int J Legal Med*, **2013**. 127(3): p. 559-72.

378. Phillips, C., A. Freire Aradas, A.K. Kriegel, M. Fondevila, O. Bulbul, C. Santos, F. Serrulla Rech, M.D. Perez Carceles, *et al.*, Eurasiaplex: a forensic SNP assay for differentiating European and South Asian ancestries. *Forensic Sci Int Genet*, **2013**. 7(3): p. 359-66.

379. Alexander, D.H., J. Novembre, and K. Lange, Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, **2009**. 19(9): p. 1655-64.

380. Hefner, J.T., Cranial nonmetric variation and estimating ancestry. *J Forensic Sci*, **2009**. 54(5): p. 985-95.

381. Bennett, D.C. and M.L. Lamoreux, The color loci of mice--a genetic century. *Pigment Cell Res*, **2003**. 16(4): p. 333-44.

382. Lamason, R.L., M.A. Mohideen, J.R. Mest, A.C. Wong, H.L. Norton, M.C. Aros, M.J. Jurynec, X. Mao, *et al.*, SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, **2005**. 310(5755): p. 1782-6.

383. Voisey, J. and A. van Daal, Agouti: from mouse to man, from skin to fat. *Pigment Cell Res*, **2002**. 15(1): p. 10-8.

384. Sulem, P., D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, K.P. Magnusson, A. Manolescu, A. Karason, *et al.*, Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet*, **2007**. 39(12): p. 1443-52.

385. Candille, S.I., D.M. Absher, S. Beleza, M. Bauchet, B. McEvoy, N.A. Garrison, J.Z. Li, R.M. Myers, *et al.*, Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. *PLoS One*, **2012**. 7(10): p. e48294.

386. Han, J., P. Kraft, H. Nan, Q. Guo, C. Chen, A. Qureshi, S.E. Hankinson, F.B. Hu, *et al.*, A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation. *PLoS Genet*, **2008**. 4(5): p. e1000074.

387. Stokowski, R.P., P.V. Pant, T. Dadd, A. Fereday, D.A. Hinds, C. Jarman, W. Filsell, R.S. Ginger, *et al.*, A genomewide association study of skin

pigmentation in a South Asian population. *Am J Hum Genet*, **2007**. 81(6): p. 1119-32.

388. Nan, H., P. Kraft, A.A. Qureshi, Q. Guo, C. Chen, S.E. Hankinson, F.B. Hu, G. Thomas, *et al.*, Genome-wide association study of tanning phenotype in a population of European ancestry. *J Invest Dermatol*, **2009**. 129(9): p. 2250-7.

389. Sulem, P., D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, M. Jakobsdottir, S. Steinberg, S.A. Gudjonsson, *et al.*, Two newly identified genetic determinants of pigmentation in Europeans. *Nat Genet*, **2008**. 40(7): p. 835-7.

390. Sturm, R.A., Molecular genetics of human pigmentation diversity. *Hum Mol Genet*, **2009**. 18(R1): p. R9-17.

391. Canfield, V.A., A. Berg, S. Peckins, S.M. Wentzel, K.C. Ang, S. Oppenheimer, and K.C. Cheng, Molecular phylogeography of a human autosomal skin color locus under natural selection. *G3: Genes Genomes*, **2013**. 3(11): p. 2059-67.

392. Basu Mallick, C., F.M. Iliescu, M. Mols, S. Hill, R. Tamang, G. Chaubey, R. Goto, S.Y. Ho, *et al.*, The light skin allele of SLC24A5 in South Asians and Europeans shares identity by descent. *PLoS Genet*, **2013**. 9(11): p. e1003912.

393. Beleza, S., N.A. Johnson, S.I. Candille, D.M. Absher, M.A. Coram, J. Lopes, J. Campos, Araujo, II, *et al.*, Genetic architecture of skin and eye color in an African-European admixed population. *PLoS Genet*, **2013**. 9(3): p. e1003372.

394. Pneuman, A., Z.M. Budimlija, T. Caragine, M. Prinz, and E. Wurmbach, Verification of eye and skin color predictors in various populations. *Leg Med (Tokyo)*, **2012**. 14(2): p. 78-83.

395. Dembinski, G.M. and C.J. Picard, Evaluation of the IrisPlex DNA-based eye color prediction assay in a United States population. *Forensic Sci Int Genet*, **2014**. 9(0): p. 111-7.

396. Wu, D.D. and Y.P. Zhang, Different level of population differentiation among human genes. *BMC Evol Biol*, **2011**. 11(1): p. 16.

397. Beleza, S., A.M. Santos, B. McEvoy, I. Alves, C. Martinho, E. Cameron, M.D. Shriver, E.J. Parra, *et al.*, The timing of pigmentation lightening in Europeans. *Mol Biol Evol*, **2013**. 30(1): p. 24-35.

398. Sankararaman, S., S. Mallick, M. Dannemann, K. Prufer, J. Kelso, S. Paabo, N. Patterson, and D. Reich, The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, **2014**. 507(7492): p. 354-7.

399. Vernot, B. and J.M. Akey, Resurrecting surviving Neandertal lineages from modern human genomes. *Science*, **2014**. 343(6174): p. 1017-21.

400. Lindsay, D.S., P.C. Jack, Jr., and M.A. Christian, Other-race face perception. *J Appl Psychol*, **1991**. 76(4): p. 587-9.

401. Luan, X., Y. Ito, Y. Zhang, and T.G. Diekwisch, Characterization of the mouse CP27 promoter and NF-Y mediated gene regulation. *Gene*, **2010**. 460(1-2): p. 8-19.

402. Samaan, G., D. Yugo, S. Rajagopalan, J. Wall, R. Donnell, D. Goldowitz, R. Gopalakrishnan, and S. Venkatachalam, Foxn3 is essential for craniofacial development in mice and a putative candidate involved in human congenital craniofacial defects. *Biochem Biophys Res Commun*, **2010**. 400(1): p. 60-5.

403. Thomas, P.S., J. Kim, S. Nunez, M. Glogauer, and V. Kaartinen, Neural crest cell-specific deletion of Rac1 results in defective cell-matrix interactions and severe craniofacial and cardiovascular malformations. *Dev Biol*, **2010**. 340(2): p. 613-25.

404. Parsons, T.E., E. Kristensen, L. Hornung, V.M. Diewert, S.K. Boyd, R.Z. German, and B. Hallgrimsson, Phenotypic variability and craniofacial dysmorphology: increased shape variance in a mouse model for cleft lip. *Journal of Anatomy*, **2008**. 212(2): p. 135-143.

405. Chai, Y. and R.E. Maxson, Jr., Recent advances in craniofacial morphogenesis. *Dev Dyn*, **2006**. 235(9): p. 2353-75.

406. Halford, M.M., J. Armes, M. Buchert, V. Meskenaite, D. Grail, M.L. Hibbs, A.F. Wilks, P.G. Farlie, *et al.*, Ryk-deficient mice exhibit craniofacial defects associated with perturbed Eph receptor crosstalk. *Nat Genet*, **2000**. 25(4): p. 414-8.

407. Clouthier, D.E., S.C. Williams, H. Yanagisawa, M. Wieduwilt, J.A. Richardson, and M. Yanagisawa, Signaling pathways crucial for craniofacial development revealed by endothelin-A receptor-deficient mice. *Dev Biol*, **2000**. 217(1): p. 10-24.

408. Nottoli, T., S. Hagopian-Donaldson, J. Zhang, A. Perkins, and T. Williams, AP-2-null cells disrupt morphogenesis of the eye, face, and limbs in chimeric mice. *Proc Natl Acad Sci U S A*, **1998**. 95(23): p. 13714-9.

409. Kaartinen, V., J.W. Voncken, C. Shuler, D. Warburton, D. Bu, N. Heisterkamp, and J. Groffen, Abnormal lung development and cleft palate in mice lacking TGF-beta 3 indicates defects of epithelial-mesenchymal interaction. *Nat Genet*, **1995**. 11(4): p. 415-21.

410. Coussens, A.K., C.R. Wilkinson, I.P. Hughes, C.P. Morris, A. van Daal, P.J. Anderson, and B.C. Powell, Unravelling the molecular control of calvarial suture fusion in children with craniosynostosis. *BMC Genomics*, **2007**. 8: p. 458.

411. Coussens, A.K., I.P. Hughes, C.R. Wilkinson, C.P. Morris, P.J. Anderson, B.C. Powell, and A. van Daal, Identification of genes differentially expressed by prematurely fused human sutures using a novel in vivo - in vitro approach. *Differentiation*, **2008**. 76(5): p. 531-45.

412. Shkoukani, M.A., M. Chen, and A. Vong, Cleft Lip - A Comprehensive Review. *Front Pediatr*, **2013**. 1: p. 53.

413. Marazita, M.L., The evolution of human genetic studies of cleft lip and cleft palate. *Annu Rev Genomics Hum Genet*, **2012**. 13(1): p. 263-83.

414. Tandon, A., N. Patterson, and D. Reich, Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. *Genet Epidemiol*, **2011**. 35(1): p. 80-3.

415. Santos, N.P., E.M. Ribeiro-Rodrigues, A.K. Ribeiro-Dos-Santos, R. Pereira, L. Gusmao, A. Amorim, J.F. Guerreiro, M.A. Zago, *et al.*, Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Hum Mutat*, **2010**. 31(2): p. 184-90.

416. Zhou, N. and L. Wang, Effective selection of informative SNPs and classification on the HapMap genotype data. *BMC Bioinformatics*, **2007**. 8(1): p. 484.

417. Royal, C.D., J. Novembre, S.M. Fullerton, D.B. Goldstein, J.C. Long, M.J. Bamshad, and A.G. Clark, Inferring genetic ancestry: opportunities, challenges, and implications. *Am J Hum Genet*, **2010**. 86(5): p. 661-73.

418. Hrdy, D.B., Analysis of hair samples of mummies from Semma South (Sudanese Nubia). *Am J Phys Anthropol*, **1978**. 49(2): p. 277-82.

419. Sturm, R.A. and M. Larsson, Genetics of human iris colour and patterns. *Pigment Cell Melanoma Res*, **2009**. 22(5): p. 544-62.

420. Fitzpatrick, T.B., The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, **1988**. 124(6): p. 869.

421. Pritchard, J.K., M. Stephens, and P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics*, **2000**. 155(2): p. 945-959.

422. Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich, Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, **2006**. 38(8): p. 904-9.

423. Patterson, N., A.L. Price, and D. Reich, Population structure and eigenanalysis. *PLoS Genet*, **2006**. 2(12): p. e190.

424. Wallenstein, S. and J. Wittes, The power of the Mantel-Haenszel test for grouped failure time data. *Biometrics*, **1993**. 49(4): p. 1077-87.

425. Johnson, A.D., R.E. Handsaker, S.L. Pulit, M.M. Nizzari, C.J. O'Donnell, and P.I. de Bakker, SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, **2008**. 24(24): p. 2938-9.

426. Boyle, A.P., E.L. Hong, M. Hariharan, Y. Cheng, M.A. Schaub, M. Kasowski, K.J. Karczewski, J. Park, *et al.*, Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*, **2012**. 22(9): p. 1790-7.

427. Tian, C., P.K. Gregersen, and M.F. Seldin, Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet*, **2008**. 17(R2): p. R143-50.

428. Wu, D.D. and Y.P. Zhang, Positive Darwinian selection in human population: A review. *Chinese Science Bulletin*, **2008**. 53(10): p. 1457-1467.

429. Online Mendelian Inheritance in Man, O. Variation in skin/hair/eye pigmentation. [cited 2014 21 January]; Available from: http://omim.org/entry/227220.

430. Vilhjalmsson, B.J. and M. Nordborg, The nature of confounding in genome-wide association studies. *Nat Rev Genet*, **2013**. 14(1): p. 1-2.

431. Barnholtz-Sloan, J.S., R. Chakraborty, T.A. Sellers, and A.G. Schwartz, Examining population stratification via individual ancestry estimates versus self-reported race. *Cancer Epidemiol Biomarkers Prev*, **2005**. 14(6): p. 1545-51.

432. Tang, H., T. Quertermous, B. Rodriguez, S.L. Kardia, X. Zhu, A. Brown, J.S. Pankow, M.A. Province, *et al.*, Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet*, **2005**. 76(2): p. 268-75.

433. Tishkoff, S.A., F.A. Reed, F.R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J.B. Hirbo, A.A. Awomoyi, *et al.*, The genetic structure and history of Africans and African Americans. *Science*, **2009**. 324(5930): p. 1035-44.

434. Kidd, K.K., W.C. Speed, A.J. Pakstis, M.R. Furtado, R. Fang, A. Madbouly, M. Maiers, M. Middha, *et al.*, Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Sci Int Genet*, **2014**. 10C(0): p. 23-32.

435. Sun, P., R. Zhang, Y. Jiang, X. Wang, J. Li, H. Lv, G. Tang, X. Guo, *et al.*, Assessing the patterns of linkage disequilibrium in genic regions of the human genome. *FEBS J*, **2011**. 278(19): p. 3748-55.

436. Haber, M., D. Gauguier, S. Youhanna, N. Patterson, P. Moorjani, L.R. Botigue, D.E. Platt, E. Matisoo-Smith, *et al.*, Genome-wide diversity in the levant reveals recent structuring by culture. *PLoS Genet*, **2013**. 9(2): p. e1003316.

437. Zalloua, P.A., Y. Xue, J. Khalife, N. Makhoul, L. Debiane, D.E. Platt, A.K. Royyuru, R.J. Herrera, *et al.*, Y-chromosomal diversity in Lebanon is structured by recent historical events. *Am J Hum Genet*, **2008**. 82(4): p. 873-82.

438. Turner, S., L.L. Armstrong, Y. Bradford, C.S. Carlson, D.C. Crawford, A.T. Crenshaw, M. de Andrade, K.F. Doheny, *et al.*, *Quality Control Procedures for Genome-Wide Association Studies*, in Current Protocols in Human Genetics. **2001**, John Wiley & Sons, Inc.

439. Pongpanich, M., P.F. Sullivan, and J.Y. Tzeng, A quality control algorithm for filtering SNPs in genome-wide association studies. *Bioinformatics*, **2010**. 26(14): p. 1731-7.

440. McIver, L.J., J.W. Fondon, 3rd, M.A. Skinner, and H.R. Garner, Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics*, **2011**. 97(4): p. 193-9.

441. Gray, R.D. and Q.D. Atkinson, Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, **2003**. 426(6965): p. 435-9.

442. Rosenberg, N.A., J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, and M.W. Feldman, Genetic structure of human populations. *Science*, **2002**. 298(5602): p. 2381-5.

443. Soejima, M. and Y. Koda, Population differences of two coding SNPs in pigmentation-related genes SLC24A5 and SLC45A2. *International Journal of Legal Medicine*, **2007**. 121(1): p. 36-39.

444. Branicki, W., U. Brudnik, J. Draus-Barini, T. Kupiec, and A. Wojas-Pelc, Association of the SLC45A2 gene with physiological human hair colour variation. *J Hum Genet*, **2008**. 53(11-12): p. 966-71.

445. Giardina, E., I. Pietrangeli, C. Martinez-Labarga, C. Martone, F. de Angelis, A. Spinella, G. De Stefano, O. Rickards, *et al.*, Haplotypes in SLC24A5 Gene as Ancestry Informative Markers in Different Populations. *Curr Genomics*, **2008**. 9(2): p. 110-4.

446. Duffy, D.L., Z.Z. Zhao, R.A. Sturm, N.K. Hayward, N.G. Martin, and G.W. Montgomery, Multiple pigmentation gene polymorphisms account for a substantial proportion of risk of cutaneous malignant melanoma. *J Invest Dermatol*, **2010**. 130(2): p. 520-8.

447. Eiberg, H., J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K.W. Kjaer, and L. Hansen, Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Hum Genet*, **2008**. 123(2): p. 177-87.

448. Mou, C., H.A. Thomason, P.M. Willan, C. Clowes, W.E. Harris, C.F. Drew, J. Dixon, M.J. Dixon, *et al.*, Enhanced ectodysplasin-A receptor (EDAR) signaling alters multiple fiber characteristics to produce the East Asian hair form. *Hum Mutat*, **2008**. 29(12): p. 1405-11.

449. Fujimoto, A., R. Kimura, J. Ohashi, K. Omi, R. Yuliwulandari, L. Batubara, M.S. Mustofa, U. Samakkarn, *et al.*, A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Hum Mol Genet*, **2008**. 17(6): p. 835-43.

450. Fantauzzo, K.A., M. Kurban, B. Levy, and A.M. Christiano, Trps1 and its target gene Sox9 regulate epithelial proliferation in the developing hair follicle and are associated with hypertrichosis. *PLoS Genet*, **2012**. 8(11): p. e1003002.

451. Takemoto, H., K. Tamai, E. Akasaka, D. Rokunohe, N. Takiyoshi, N. Umegaki, K. Nakajima, T. Aizu, *et al.*, Relation between the expression levels of the POU transcription factors Skn-1a and Skn-1n and keratinocyte differentiation. *J Dermatol Sci*, **2010**. 60(3): p. 203-5.

452. Donnelly, M.P., P. Paschou, E. Grigorenko, D. Gurwitz, C. Barta, R.B. Lu, O.V. Zhukova, J.J. Kim, *et al.*, A global view of the OCA2-HERC2 region and pigmentation. *Hum Genet*, **2012**. 131(5): p. 683-96.

453. Cappuccio, G., R. Genesio, V. Ronga, A. Casertano, A. Izzo, M.P. Riccio, C. Bravaccio, M.C. Salerno, *et al.*, Complex chromosomal rearrangements causing Langer-Giedion syndrome atypical phenotype: Genotype-phenotype correlation and literature review. *Am J Med Genet A*, **2014**. 164(3): p. 753-9.

454. Beaumont, K.A., S.N. Shekar, A.L. Cook, D.L. Duffy, and R.A. Sturm, Red hair is the null phenotype of MC1R. *Hum. Mutat.*, **2008**. 29: p. e88-e94.

455. Girotto, G., D. Vuckovic, A. Buniello, B. Lorente-Canovas, M. Lewis, P. Gasparini, and K.P. Steel, Expression and replication studies to identify new candidate genes involved in normal hearing function. *PLoS One*, **2014**. 9(1): p. e85352.

456. Fitch, K.R., K.A. McGowan, C.D. van Raamsdonk, H. Fuchs, D. Lee, A. Puech, Y. Herault, D.W. Threadgill, *et al.*, Genetics of dark skin in mice. *Genes Dev*, **2003**. 17(2): p. 214-28.

457. Terunuma, A. and J. Vogel, Gene sets for detection of ultraviolet a exposure and methods of use thereof. **2013**, *Google Patents*.

458. Lokody, I., Complex traits: Non-coding polymorphism in IRF4 reveals function. *Nature Rev Genet*, **2013**. 15(1): p. 5-5.

459. Praetorius, C., C. Grill, S.N. Stacey, A.M. Metcalf, D.U. Gorkin, K.C. Robinson, E. Van Otterloo, R.S. Kim, *et al.*, A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. *Cell*, **2013**. 155(5): p. 1022-33.

460. Chakravarty, M.M., R. Aleong, G. Leonard, M. Perron, G.B. Pike, L. Richer, S. Veillette, Z. Pausova, *et al.*, Automated analysis of craniofacial morphology using magnetic resonance images. *PLoS One*, **2011**. 6(5): p. e20241.

461. Fernandez-Riveiro, P., E. Smyth-Chamosa, D. Suarez-Quintanilla, and M. Suarez-Cunqueiro, Angular photogrammetric analysis of the soft tissue facial profile. *Eur J Orthod*, **2003**. 25(4): p. 393-9.

462. Weston, E.M., A.E. Friday, and P. Lio, Biometric evidence that sexual selection has shaped the hominin face. *PLoS One*, **2007**. 2(8): p. e710.

463. Galdzicka, M., S. Patnala, M.G. Hirshman, J.F. Cai, H. Nitowsky, J. A Egeland, and E.I. Ginns, A new gene, EVC2, is mutated in Ellis–van Creveld syndrome. *Mol Genet Metab*, **2002**. 77(4): p. 291-295.

464. Novelli, G., A. Muchir, F. Sangiuolo, A. Helbling-Leclerc, M.R. D'Apice, C. Massart, F. Capon, P. Sbraccia, *et al.*, Mandibuloacral dysplasia is caused by a mutation in LMNA-encoding lamin A/C. *Am J Hum Genet*, **2002**. 71(2): p. 426-31.

465. Girirajan, S., S. Williams, J. Garbern, N. Nowak, E. Hatchwell, and S. Elsea, 17p11.2p12 triplication and del(17)q11.2q12 in a severely affected child with dup(17)p11.2p12 syndrome. *Clin Genet*, **2007**. 72(1): p. 47-58.

466. Allache, R., P. De Marco, E. Merello, V. Capra, and Z. Kibar, Role of the planar cell polarity gene CELSR1 in neural tube defects and caudal agenesis. *Birth Defects Res A Clin Mol Teratol*, **2012**. 94(3): p. 176-81.

467. Meng, Q., C. Jin, Y. Chen, J. Chen, M. Medvedovic, and Y. Xia, Expression of signaling components in embryonic eyelid epithelium. *PLoS One*, **2014**. 9(2): p. e87038.

468. Psychiatric, G.C.B.D.W.G., Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet*, **2011**. 43(10): p. 977-83.

469. Hirschhorn, J.N. and G. Lettre, Progress in genome-wide association studies of human height. *Horm Res*, **2009**. 71 Suppl 2: p. 5-13.

470. Browning, S.R. and B.L. Browning, Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*, **2007**. 81(5): p. 1084-97.

471. Pereira, R., C. Phillips, N. Pinto, C. Santos, S.E. dos Santos, A. Amorim, A. Carracedo, and L. Gusmao, Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing. *PLoS One*, **2012**. 7(1): p. e29684.

Supplemental materials

Table S1. A list of loss-of-function mutations that cause craniofacial defects in mice

Gene	Protein name	Major defects	Primary abnormality	Human orthologue	Human disorder (if known)
Acvr2	Activin receptor IIA	Meckel's cartilage/mandibular hypoplasia, cleft palate, defective eyelid closure (22%)	Altered signalling during early mandible development	ACVR2	–
Alx3/4	Aristaless 4	Reduced size of parietal bone	Delay in ossification	ALX4	Parietal foramina
Apaf1	Apoptotic protease-activating factor 1	Midline facial cleft, absence of skull vault, vomer, ethmoid,rostral exencephaly, cleft palate	Participates in apoptosis	APAF1	–
Apob	Apolipoprotein B	Exencephaly	–	APOB	Hypobetalipoproteinemia
Bapx1	Bagpipe homeobox 1	Reduced or absent supraoccipital,exoccipital, basioccipital,basisphenoid bones	Early marker of chondrogenesis	BAPX1	–
Bmp1	Bone morpho-genetic protein 1	Reduced size of frontal, parietal and interparietal bones	Collagen-processing defect	BMP1	–
Bmp4	Bone morpho-genetic protein 4	Haploinsufficiency: short, bent frontal and nasal bones (12%), eye defects (35%).	Required for mesoderm formation at gastrulation	BMP4	–
Bmp5	Bone morpho-genetic protein 5	Small external ears	–	BMP5	–
Bmp7	Bone morpho-genetic protein 7	Anophthalmia	At or after lens induction	BMP7	–
Cart1	Cartilage homeoprotein 1	Failure of neural-tube closure	Deficit of forebrain mesenchyme caused byexcessive cell death	CART1	–
Chrd	Chordin	Defective inner and outer ear development	–	CHRD	–
Chuk1	Conserved helix– loop–helix ubiquitouskinase (IKKα)	Small, missing, maldevelopedskull bones	Intact IKK activation	CHUK	–
Chx10	ceh-10 homeo-domain homologue	Microphthalmia, optic nerve aplasia, cataract	Impaired retinal progenitor proliferation and bipolar-cell differentiation	CHX10	Microphthalmia, cataracts
Col11a1	Procollagen, type XI, α1	Short mandible, cleft palate	Disorganized extracellularmatrix	COL11A1	Stickler syndrome
Col2a1	Procollagen,	Bulging forehead, short snout, cleft	Defective endochondral	COL2A1	Stickler

	type II, α1	palate	bone formation		syndrome, Kniest dysplasia,achondrogenesis type II
Crkol	v-crk avian sarcoma CT10 oncogenehomologue-like	Short, broad skull	Post-migratory defect ofneural crest	CRKL	–
Crtl1	Cartilage linkprotein 1	Short skull base with reduced ethmoid, sphenoid body, occipital bone, Meckel'scartilage, middle ear bones	Required to form cartilage proteoglycan aggregates	CRTL1	–
Dhcr7	7-dehydrochol-esterol reductase	Cleft palate (8%))	Palatal shelves elevate but fail to fuse	DHCR7	Smith–Lemli–Opitzsyndrome
Dlx1	Distal-less homeobox	Abnormal ala temporalis, minorabnormalities of some other bones, small cleft palate in 10%of cases	-	DLX1	–
Dlx2	Distal-less homeobox 2	Abnormal derivatives of first and second arches, 80% cleft palate; more severe if Dlx1-/-	Respecification of fate of asubset of neural-crest cells	DLX2	–
Dlx5	Distal-less homeobox 5	28% exencephaly, hypomineralized – parietal/ interparietal bones,abnormal nasal and otic capsules,branchial arches, 88% cleft palate,malformed teeth		DLX5	–
Dnmt3b	DNA methyltrans-ferase 3b	Neural-tube defect	Defective methylation	DNMT3B	ICF syndrome
Ece1	Endothelin-converting enzyme	Reduced mandible, alisphenoid,squamosal, palatine bones,absent tympanic ring, Meckel's cartilage	Abnormal developmentof subsets of cephalicneural-crest tissue	ECE1	Hirschsprung disease,cardiac defects dysmorphism
Edn1	Endothelin 1	Reduced mandible, zygomatic,temporal bones, absent auditoryossicles and tympanic ring	Abnormal developmentof subsets of cephalicneural-crest tissue	EDN1	–
Ednra	Endothelin receptor type A	Reduced mandible, alisphenoid,squamosal, palatine bones, absent tympanic ring, Meckel's cartilage	Abnormal developmentof subsets of cephalicneural-crest tissue	EDNRA	–
Egfr	Epidermal growth factor receptor	Short mandible, 5% cleft palate	Abnormal morphogenesisof Meckel's cartilage	EGFR	–
Eya1	Eyes absent 1	Homozygote: absent/malformed		EYA1	Branchio-oto-

		Arrest of inner ear at otic auricles, meatus, malleus, inner ear, vesicle stage with mandible, maxilla, reduced skull increased apoptosis ossification. Heterozygote: hearing loss due to abnormal sound conduction in middle ear			renal syndrome
Folbp1	Folate-bindingprotein 1	Neural-tube defect	Thin neuroepithelium with absent forebrain and opticvesicles	FOLR1	–
Foxa2	Forkhead box a2 (HNF3β)	Loss of structures anterior to hindbrain	Required for node and notochord formation	HNF3B	–
Foxc2	Forkhead box C2 (Mfh1)		Defects of several skull bones	FOXC2	Lymphoedema–distichiasis syndrome
Foxe1	Forkhead box E1 (thyroid transcription factor 2)		Defective migration ofthyroid precursor cells, palatal shelves movecorrectly but do not fuse	FOXE1	Thyroid agenesis, cleft palate, choanal atresia
Fst	Follistatin	Cleft palate or absent hard palate (16–55%), delayed/absent lower incisors	Defective chondrogenesis/ bone formation	FST	–
Gabrb3	γ-amino butyricacid receptor,subunit β3	Cleft palate		GABRB3	–
Gad1	Glutamic acid decarboxylase 1(Gad67)	Cleft palate	Palatal shelves elevate but fail to fuse	GAD1	–
Gja1	Gap junctionmembrane channel protein α1 (Connexin43)	Delayed membranous andendochondral ossification of skull	Impaired neural-crest migration and osteoblastdysfunction	GJA1	–
Gli2	GLI-Kruppel familymember 2	Deficient medial ossification of frontal and parietal bones,absent upper/lower incisors,cleft palate	Skeletal development	GLI2	–
Gli3	GLI-Kruppel familymember 3	Neural-tube defect, largemaxilla and premaxilla	Defective neural-tube closure and skeletogenesis	GLI3	Greig, Pallister Hall syndromes
Gp330	Glycoprotein 330 (megalin)	Holoprosencephaly	Reduced telencephalicvesicle; mediates endocytosis of lipoproteins andother	LRP2	–

			macromolecules		
Gsc	Goosecoid	Hypoplastic lower mandible,aplastic nasal cavity, capsule, inner ear	Expression in neural-crest derivates (mandibulararch) after E10.5	GSC	–
Hes1	Hairy and enhancerof split 1	Failure of neural-tube closure	Premature neurogenesis	HRY	–
Hesx1	Homeobox geneexpressed in ES cells	Reduction of telencephalicvesicles, eyes, olfactoryplacodes, frontonasal mass	First expressed in anterior ventral endoderm; defectiveinduction of prospective prosencephalic neuro-ectoderm	HESX1	Septo-optic dysplasia
Hhex	Haematopoietica llyexpressed homeobox	Variable anterior truncations	Required for specification of axial mesendoderm	HHEX	–
Hic1	Hypermethylate din cancer 1	Variable, 50% with gross defects including acrania, exencephaly,cleft palate	–	HIC1	–
Hoxa1	Homeobox A1	Delayed neural-tube closure, abnormal basioccipital, ex-occipital, interparietal bones,inner ear	Abnormal patterning of hindbrain rhombomeres and associated neural crest	HOXA1	–
Hoxa2	Homeobox A2	Duplication of proximal first arch elements (malleus, incus, gonial, squamosal, pterygoidbones), 80% cleft palate	Homeotic respecification of second to first arch identity	HOXA2	–
Hoxa3	Homeobox A3	Mild shortening of mandible	Abnormal developmentof pharyngeal arches,especially 3 and 4	HOXA3	–
Hspg2	Perlecan	40% die at E10.5 with defective head development; 6%exencephaly, rest short-domed skull, cleft palate, wide sutures	Inhibition of chondrocyte proliferation	HSPG2	Schwartz–Jampelsyndrome, dyssegmented dysplasia
Inhba	Activin βA	Cleft seconary palate (variable %),absent lower incisors	Defective chondrogenesis/ bone formation	INHBA	
Itgav	Integrin αv	Cleft palate in 20% survivors	Delayed growth in head structures	ITGAV	
Jag2	Jagged 2	Cleft secondary palate	Failure of palatal shelves to elevate	JAG2	
Klf2	Kruppel-like factor2 (lung)	Malformed lower jaw	Secondary to haemato-logical defect	KLF2	

Lef1	Lymphoid-enhancerbinding factor 1	Loss of teeth	–	LEF1	–
Lhx1	LIM homeobox protein 1	Absent node at E7.5, lack ofhead structures anterior to otic vesicle	First expressed in anterioventral endoderm;failure of brain specificationjust anterior to rhomomere 3	LHX1	–
Lhx8	LIM homeobox protein 8	Isolated cleft secondary palate	Palatal shelves form and elevate, but fail to fuse	–	–
Lmx1b	LIM homeobox transcription factor 1β	Absent fontanelles and supra-occipital bone, small interparietalbone, partial cranial suture fusion	–	LMX1B	Nail–patella syndrome
Macs	Myristoylatedalanine-rich proteinkinase C substrate	Exencephaly (25%)	–	MACS	–
Madh2	MAD homologue 2	Haploinsufficiency: absent or hypoplastic mandible, absenteye (E6.0, lethal)	Required for organization of primitive germ layersbefore gastrulation	MADH2	–
Mmp14	Matrix metallo-proteinase 14 (membrane-inserted)	Domed head, prominent cranial sutures (birth)	Generalized ossification defect	MMP14	–
Msx1	Msh-like homeobox 1	Short mandible, absent incisors,cleft secondary palate, minorabnormality of skull vault	Differentiation of pharyngeal arch ectomesenchyme into bones and teeth,impaired development ofpalatal shelves	MSX1	Selective tooth agenesis, cleft lip/palate
Msx2	Msh-like homeobox 2	Frontal bone defect, small inter-parietal and supraoccipitalbones	Reduced progenitors in osteogenic fronts	MSX2	Craniosynostosis, Boston type; parietalforamina
Nog	Noggin	Exencephaly	Required for neural-tube patterning	NOG	Proximal symphalangism, multiple synostosessyndrome 1
Otx2	Orthodenticle homologue 2	Homozygote: reduced embryonic ectoderm at E7.5,absent fore- and midbrain regions at E9.5 (E7.5).Heterozygote: normal (16%),through to acrania (3%),most	First expressed in anterio-ventral endoderm; primaryrole in specification of head	OTX2	–

		with micro/agnathiaand micro/anophthalmia			
Pax2	Paired box 2	Exencephaly, failure to close optic fissure, abnormal otic vesicle	Regulator of pattern in eye and ear	PAX2	Renal coloboma syndrome
Pax3	Paired box 3	Neural-tube defect, deficiencyof melanocytes, Schwann cells,dorsal root ganglia	Abnormal neural-crest development	PAX3	Waardenburg syndrome type 1
Pax6	Paired box 6	Homozygote: absent eyes andnasal cavities. Heterozygote: small eyes, iris hypoplasia	Aberrant lens and nasal placode formation, delayedclosure of optic fissure	PAX6	Aniridia
Pax7	Paired box 7	Short maxilla, hypoplastic nasalstructures	–	PAX7	–
Pax9	Paired box 9	Cleft secondary palate, absentteeth	–	PAX9	Oligodontia
Pcsk6	Proprotein convertase subtilisin/kexin type 6	Cyclopia, anterior truncationswith absence of telencephalon, nasal capsule, upper and lowerjaw, laterality effects	Impaired processing of BMPs	PACE4	–
Pdgfra	Platelet-derived growth factor receptor, α-poly-peptide	Cleft face, bleb over neural tube	Failure of subset of non-neuronal neural-crest cells to migrate	PDGFRA	–
Pitx1	Paired-like homeodomain transcriptionfactor 1	Very short mandible, normal teeth	Required for bone development	PITX1	–
Pitx2	Paired-like homeodomain transcriptionfactor 2	Cleft palate, abnormal maxillaand mandible with arrested tooth development, malformedeyes	Abnormal Fgf8 andBmp4 signalling duringmorphogenesis	PITX2	Rieger syndrome
Prrx1	Paired related homeobox 1	Cleft secondary palate, absentsquamosal, zygomatic, tympanicring, hypoplastic mandible	Defective growth of mandibular arch components; normalneural-crest migration	PMX1	–
Raldh2	Retinaldehydede hydrogenase 2	Short frontonasal region, neural-tube defect	Synthetic deficiency ofretinoic acid; rescued bymaternal RA administration	RALDH2	–
Rax	Retina and anterior neural fold homeobox	Anophthalmia, absent forebrain, reduced midbrain	Failure of optic cup formation	RX	Severe congenital microphthalmia

Runx2	Runt-related transcription factor 2	Homozygote: defectiveossification of all bones (E17.5).Heterozygote: defectiveossification of skull vault, clavicle,hypoplastic hyoid, pubic, ischial,wide xiphoid process	Required for osteoblast differentiation	RUNX2	Cleidocranial dysplasia
Ryk	Receptor-like tyrosine kinase	Cleft secondary palate (88%),short snout, small, rounded cranial vault	Defective extrapalatalmorphogenesis	RYK	–
Shh	Sonic hedgehog	Central optic vesicle, underdeveloped mid- and forebrain	Maintenance of notochord	SHH	Holoprosencephaly
Sil	Tal1 interrupting locus	Cyclopia and neural-tube defect; lethal by	Possible block in Shh-mediated signalling	SIL	–
Ski	Sloan–Ketteringviral oncogenehomologue	Failure of neural-tube closure, absent presphenoid, small mandible, basioccipital,basisphenoid	Excessive apoptosis incranial mesenchymeand neural tube	SKI	–
Tbx1	T-box 1	Abnormal mandible, ears; cleft palate	Absent second and third TBX1 branchial arches, otic vesicles		DiGeorge syndrome
Tcfap2a	Transcription factor	Exencephaly, reduced maxilla and mandible	Neural-crest migration normal; excess neuroepithelial proliferation and cell death	TFAP2A	–
Tcof1	Treacher Collins–Franceschetti syndrome 1, homologue	Exencephaly, anophthalmia, nasal agenesis, abnormalmaxilla	Neural-crest apoptosis	TCOF1	Treacher Collins–Franceschetti syndrome
Tgfb2	Transforming growth factor β2	Cleft palate (23%), reduced size and ossification of frontal interparietal, parietal, squamosal	Neural crest and skeletogenesis	TGFB2	–
Tgfb3	Transforming growth factor β3	Isolated cleft secondary palate	Palatal shelves apposebut fail to fuse	TGFB3	–
Trp53	Transformation-related protein 53(p53)	8–16% exencephaly	–	TP53	Li–Fraumeni syndrome
Trp63	Transformation-related protein 63 (p63)	Small mandible	Required for epithelial development	TP63	Ectrodactyly, ectodermal dysplasia,clefting
Twist	Twist	Failure of neural-tube closure, abnormal branchial arches	Mesenchymal expression required for neural-tube	TWIST	Saethre–Chotzen syndrome

			closure and neural-crest		
Vax1	Ventral anterior homeobox 1	contribution to branchial			–
Vcl	Vinculin	arches			–
Wnt5a	Wingless-related MMTV integrationSite 5A	Lobar holoprosencephaly, fused maxillary incisors, cleft palate	Required for optic nerve and forebrain development; also first branchial arch	VAX1	–
Zfx1a	Zinc-fingerhomeobox 1a (δEF1)	Neural-tube defect	Failure of closure	VCL	–
		Short maxilla and mandible	Reduced amount of mesoderm leading toreduced outgrowth	WNT5A	–
		Neural-tube defect (1%), shortdistal maxilla and mandible,hyperplastic Meckel's cartilage,cleft palate	Neural crest and skeletogenesis	TCF8	–

Table S2: Common human craniofacial disorders with known single gene mutations.

Disorder	Gene	Major craniofacial abnormalities
Parietal foramina 2	**ALX4**	Parietal foramina, cranium bifidum
Gardner syndrome	**APC**	Jaw cysts
Chondrodysplasia punctata, X-linked recessive	**ARSE**	Nasal hypoplasia
Marshall syndrome	**COL11A1**	Flat midface, cleft palate, hearing loss
Stickler syndrome, type II		Cleft palate, micrognathia, Myopia, beaded vitreous hearing loss
Otospondylomega-epiphyseal dysplasia syndrome	**COL11A2**	Cleft palate, deafness
Stickler syndrome, type III, Weissenbacher–Zweymuller syndrome		Cleft palate, micrognathia, hearing loss
Knobloch syndrome	**COL18A1**	Occipital encephalocele
Kniest dysplasia	**COL2A1**	Cleft palate
Spondyloepimetaphyseal dysplasia, Strudwick type Stickler syndrome, type I		Cleft palate, micrognathia, Myopia, retinal hearing lossdetachment
Pycnodysostosis	**CTSK**	Hypoplastic mandible, delayed eruption of teeth, wide cranial sutures
Smith–Lemli–Opitz syndrome	**DHCR7**	Cleft palate, holoprosencephaly
Trichodentoosseous syndrome	**DLX3**	Macrocephaly, skull sclerosis, delayed dental eruption
Dentinogenesis imperfecta	**DSPP**	Abnormal dentin production and mineralization, sensorineural hearing loss

Atelosteogenesis dysplasia I (Neonatal osseus type II)	**DTDST**	Cleft palate
Diastrophic dysplasia		Malformed external ears, cleft palate
Ectodermal dysplasia, anhydrotic	**ED1**	Absent or pointed teeth
Ellis–van Creveld syndrome	**EVC**	Lip-tie, hypodontia
Weyers acrofacial dysostosis		Conical teeth, abnormal mandible
Branchio-oto-renal dysplasia	**EYA1**	Branchial fistulas, malformed ears, cochlear malformation
Shprintzen–Goldberg syndrome	**FBN1**	Craniosynostosis
Pfeiffer syndrome	**FGFR1**	Craniosynostosis
Apert syndrome **Beare–Stevenson cutis gyrata syndrome** **Crouzon syndrome, Jackson–Weiss syndrome**	**FGFR2**	Craniosynostosis
Achondroplasia	**FGFR3**	Macrocephaly
Crouzon syndrome with acanthosis nigricans		Craniosynostosis, cementomas of jaw
Muenke syndrome		Coronal synostosis
Thanatophoric dysplasia, types I and II		Severe craniosynostosis
Lymphoedema–distichiasis syndrome	**FOXC2**	Cleft palate

Hypothyroidism, athyroidal, with spiky hair and cleft palate	**FOXE1**	Cleft palate
Greig cephalopoly-syndactyly	**GLI3**	Macrocephaly
Pallister–Hall syndrome		Cleft palate, microtia
Simpson–Golabi–Behmel syndrome, type 1	**GPC3**	Macrocephaly, cleft palate
Dyssegmental dysplasia, Silverman–Handmaker type	**HSPG2**	Cleft palate, small jaw, flat face, encephalocele
Hyaluronidase deficiency	**HYAL1**	Submucous cleft palate
Opitz G syndrome, type I	**MID1**	Hypertelorism, cleft lip/palate
Hypodontia with orofacial cleft	**MSX1**	Cleft lip/palate, hypodontia
Craniosynostosis, type 2 (Boston)	**MSX2**	Variable craniosynostosis
Parietal foramina 1		Parietal foramina
Oral–facial–digital syndome	**OFD1**	Midline clefts, tongue hamartomas, dystopia canthorum
Craniofacial–deafness–hand syndrome	**PAX3**	Hypertelorism, hypoplastic nose, hearing loss
Waardenburg syndrome, types I and III		Dystopia canthorum
Aniridia	**PAX6**	Aniridia, Peters anomaly, foveal hypoplasia, keratitis anophthalmia (homozygous)
Oligodontia	**PAX9**	Oligodontia

Rieger syndrome, type 1	**PITX2**	Hypodontia
Basal cell nevus syndrome	**PTCH**	Jaw cysts, macrocephaly
Bannayan–Zonana syndrome	**PTEN**	Macrocephaly
Metaphyseal chondrodysplasia, Murk Jansen	**PTHR**	Cranial sclerosis, choanal stenosis
Cleft lip/palate ectodermal dysplasia syndrome, Zlotogora– Ogur syndrome	**PVRL1**	Delayed membraneous skull ossification, dental abnormalities
Robinow syndrome, autosomal recessive	**ROR2**	Macrocephaly, facial cleft
Cleidocranial dysplasia	**RUNX2**	Cleft lip/palate
Dental anomalies, isolated		Delayed eruption of permanent teeth
Townes–Brocks syndrome	**SALL1**	Dysplastic ears, preauricular pits/tags, hearing loss
Holoprosencephaly-3	**SHH**	Holoprosencephaly, cyclopia, proboscis
Hirschsprung disease with microcephaly, mental retardation, distinct facial features	**SIP1**	Hypertelorism, microcephaly
Holoprosencephaly-2	**SIX3**	Holoprosencephaly, cleft lip/palate
Sclerosteosis	**SOST**	Cortical hyperostosis
Van Buchem disease		Osteosclerosis of skull, mandible
Campomelic dysplasia with autosomal sex reversal	**SOX9**	Cleft palate
DiGeorge syndrome	**TBX1**	Cleft palate
Treacher Collins mandibulofacial dysostosis	**TCOF1**	Malar hypoplasia, ear anomalies, cleft palate
Holoprosencephaly-4	**TGIF**	Lobar or semilobar holoprosencephaly, midline

		cleft lip/palate
Ectrodactyly, ectodermal dysplasia, and cleft lip/palate syndrome 3	**TP63**	Cleft lip/palate
Saethre–Chotzen syndrome	**TWIST**	Craniosynostosis
Holoprosencephaly-5	**ZIC2**	Alobar or semilobar holoprosencephaly, mild facial anomaly

BOND UNIVERSITY

STATEMENT OF CONSENT

I, (Name, please print) _____

agree to participate in this study.

I have read the attached explanatory information.

I understand that this study will be of no direct benefit to me.

I understand that participation in this study is voluntary and that I am free to withdraw at any time without penalty or comment.

☐ I am happy for a blood/saliva sample to be taken.

☐ I am happy for an image of my face to be taken

☐ I am happy for my sample to be used for other studies in the future.

☐ I am happy to complete the questionnaire.

Please tick the appropriate boxes.

If you are under 18, a parent/guardian signature is required.

I understand that my anonymity will be strictly preserved and photographs will be used only to classify my facial features and colouring.

Signature of Participant _____

Date _____

Signature of parent/guardian_____

Date _____

BOND UNIVERSITY

EXPLANATORY STATEMENT

Project Title: Inheritance of complex physical characteristics such as hair colour

Chief Investigator: Dr Angela van Daal

Faculty of Health Sciences and Medicine, Bond University.

Gold Coast, Queensland 4229

Ph: 07 5595 4433

e-mail: avandaal@staff.bond.edu.au

Bond University is committed to research in various forms. In order to conduct some research projects biological samples from individuals are required.

This project aims to investigate the inheritance of complex physical characteristics (eg hair colour, height) by analysing genes identified as associated with the particular physical trait. To carry out this research biological samples from individuals of families or groups with a high incidence of the particular physical characteristic are required (eg families containing many individuals with red hair).

Your involvement will be limited to a buccal swab sample (swab from inner cheek) or blood sample. A buccal swab is collected by simply rubbing a cotton bud (or something similar) firmly against the inside of your cheek. This is allowed to dry before placing it in the plastic sleeve. Alternatively or additionally a blood sample will be collected.

If you do not wish your sample to be used for these purposes, then you must not volunteer for this study.
Before signing this consent form you should be aware that your participation is voluntary. If you decide to take part you may also discontinue your participation at any time without comment or penalty.

If you are under 18, a parent/guardian signature is required

The results of this research may be published at a future date. Your anonymity will be preserved.

You may contact the Chief Investigator about any matter of concern (see address and contact details above) should you wish to raise any concern. If you have any concerns about the ethical conduct of this research, please contact the Bond University Human Research Ethics Committee located at Level 2, Central Building, Bond University QLD 4229 Phone: 5595 4194 Fax: 5595 4122 E-mail: buhrec@bond.edu.au

Supplemental document S4- Questionnaire

BOND UNIVERSITY

QUESTIONNAIRE

Name: _____

Date of Birth: _____ Email/Phone: _____

Please answer the following questions for your appearance as at about 20 years of age by circling the most appropriate category.

Ancestry	Caucasian	Asian	Indian	Indigenous Australian	African	Other
Ancestry Details						
Face/Head measurements	**v-gn (craniofacial height)**		**eu-eu (head width)**		**g-op (head length)**	
HAIR						
Natural Colour & Shade:	**RED**	Auburn	Orange	Carrot	Strawberry	
	BLONDE	Light Blonde		Dark Blonde		
	BROWN	Light Brown	Medium Brown	Dark Brown	Brown with red	
	BLACK	Light Black		Dark Black		
***Significant* hair colour change from child to adult?**			No		Yes	
			got much darker		got lighter	
Curliness	VERY CURLY	CURLY	HAS BODY		STRAIGHT	
Beard	Red	Blonde	Brown	Black	Mixture	
Eyebrows	Red	Blonde	Brown	Black	Mixture	
EYES	Eyelid		single		double	
Colour & Shade	BLUE	Light Blue	Medium Blue		Dark Blue	
	GREEN	Light Green	Medium Green		Dark Green	
	GRAY	Light Gray	Medium Gray		Dark Gray	
	HAZEL	Light Hazel	Medium Hazel		Dark Hazel	
	BROWN	Light Brown	Medium Brown		Dark Brown	
Skin Colour & Freckling	**FAIR**	Extensive Freckling	Medium Freckling	Light Freckling	No Freckling	
	AVERAGE	Extensive Freckling	Medium Freckling	Light Freckling	No Freckling	
	DARK	Olive	Light Brown	Dark Brown	Black	
Ears	attached		non - attached			
Height	Your height (cm or feet, inches)					
Weight	Your current weight (kg or stones, pounds)					
Weight	Your weight at 20-25 years of age (kgs, stones, pounds)					
Have you had any severe facial injuries or undergone facial surgery? If yes, please provide general details						